

Honesty and denial at NASA

The world's leading space agency is suffering not only from managerial dysfunction, but also from a failure to address strategic issues. NASA and its stakeholders need to face up to the challenges ahead.

The investigative board charged with finding the causes of February's explosion of the space shuttle Columbia turned in about as good a report last week as anyone could have asked. In clear, direct language, the panel, which was led by retired Navy admiral Harold Gehman, spelled out what has gone wrong — technically, politically, even sociologically — with the space-shuttle programme since its inception almost 30 years ago. The accident had several causes, but one of the board's conclusions looms larger than the rest: NASA has become “an agency trying to do too much with too little”.

After the report was released, NASA administrator Sean O'Keefe said all the right, contrite words. “We get it,” he assured reporters. The question is whether anyone else in Washington does, and how long the willingness to change will last. Missing from politicians' comments about the report was an acknowledgement that they, too, had been indicted, along with the shuttle-programme managers who became lax about safety.

Gehman's report accurately describes the dysfunctional relationship that has developed between the space agency and those in the White House and Congress who fund it. NASA no longer has anything like the budget or the political mandate it enjoyed in its glory days. But rather than scale back its wish list, “NASA continued to push an ambitious agenda of space science and exploration, including a costly Space Station,” the report says. The agency got into the dangerous habit of promising what it could not deliver. Meanwhile, NASA's overseers in the White House and Congress, who surely suspected that the agency was stretched too thin, suppressed their own doubts, just as shuttle engineers ignored worries about falling foam.

Everyone involved in funding the space programme has been in a state of denial. And now it is time for honesty. First, admit that space travel is inherently risky; worse, that it is risky and expensive. More funding will help — in fact, it's essential — but it cannot guarantee safety. So if we want to continue sending people into space, we should expect to keep paying in both money and lives.

Money pit

Nature has argued that science alone can't justify the astronaut programme, but that a desire for adventure and an expansion of the human spirit certainly could. To be truly inspiring, though, NASA will need to venture once again beyond Earth's orbit — an expensive and technically daunting undertaking.

In the meantime, there's the International Space Station to finish. Few Americans, even space aficionados, have much interest in that project, which has a history every bit as troubling as the shuttle's. Before President Ronald Reagan approved the station in 1984, key members of his cabinet, including his secretaries of defence and treasury, opposed it as a money pit of limited utility. White House science adviser George Keyworth was among those arguing at the time for something more grand, such as a return to the Moon. But NASA administrator James Beggs and his staff calculated that the (then) \$8 billion station was all they could get away with politically, and convinced Reagan to endorse it.

Many advocates of a bold space programme would love to have that moment back. In the years since, it has often seemed that the space station, with its constant political and budgetary woes (the price tag is now \$24 billion, not counting the cost of launch and operations), is preventing rather than enabling more ambitious space exploration. Yet the United States and its partners — Europe, Russia, Japan and Canada — will finish the project they started, primarily because it would be too embarrassing to do otherwise.

What then? Gehman's panel has called for a national debate on America's goals in space. It's difficult to imagine the United States turning its back completely on human space exploration, but any ventures bold enough to be worth the risk will require substantially more investment than the country has been willing to make since the days of Apollo.

Making a difference

Among those awaiting a solution to this dilemma are the managers of NASA's science programmes, who have their own financial woes and schedule pressures — witness the recent debate over whether to extend the life of the Hubble Space Telescope after its funding runs out in 2010. Science missions are not immune to the cultural problems that plagued the shuttle programme. Managers of the two failed Mars missions of 1999, for example, were unrealistic about how much they could accomplish on a tight schedule and skimpy budget. Gehman's analysis of shuttle workers could just as well have applied to them: “No one at NASA wants to be the one to stand up and say, ‘We can't make that date.’”

Virtually every science project at NASA is running perilously close to its cost and schedule margins. Yet the agency keeps proposing new multibillion-dollar initiatives, such as the Prometheus nuclear propulsion programme, which would enable ambitious missions to explore the outer Solar System, and the Beyond Einstein missions to study black holes and dark matter. If NASA presses on with human space flight but does not receive substantially more money, there is little hope that these new science programmes will survive.

Cancelling new and exciting science initiatives to pay for a half-hearted astronaut programme that intends merely to keep circling the Earth for another decade or two would be an unacceptable outcome. So the point bears repeating: the United States should either be prepared to invest more money in human space flight, in addition to NASA's current budget, or walk away from the challenge.

The window of opportunity for honest debate on this topic may soon close. Gehman's report was released in the doldrums of August, a time when few lawmakers are in Washington. Now that the verdict has been heard, the press and public will shift their attention elsewhere. Attempts at reform could easily get lost in the fine print of appropriations mark-ups and budget negotiations, to which few people outside the Washington establishment pay attention. Last week Gehman warned about the risk of “backsliding” to old habits, and issued his own call to action: The loss of the Columbia astronauts' lives “had better make a difference, or they and we will have wasted our time”.





Scrapped paper
Results on cancer vaccine retracted after three years
p4



Under arrest
US scientists rally behind beleaguered plague researcher
p5



Ivory tower
Lion man carves a niche for itself as oldest statue
p7



An ill wind
Forest fires present hot prospect for fungus researchers
p8

Experts fear network paralysis as computer worms blast Internet

Declan Butler, Paris

Blaster, Welchia, SoBig — the late summer blitz of computer worms was the last thing the world's universities and researchers needed as they prepared for the start of term. And worse could be on the way, computer scientists say.

Universities were seriously disrupted by the SoBig.E worm in late July — which copied itself into reams of e-mails that it sent out from infected machines — and the Blaster worm, which struck on 11 August causing computers to restart constantly. Next came the Welchia worm, which tried to fix computers infected with Blaster but ended up causing its own problems. Then, on 18 August, came the SoBig.F variant, the most voluminous e-mail worm in Internet history.

"The volume of traffic created by the worms gorged many university networks, often grinding them to a halt," says Theresa Rowe, a security official at Oakland University in Rochester, Michigan.

Despite the widespread disruption, many computer professionals say that the attacks could have been far worse. SoBig's worm-laden messages may have accounted for one in every 17 e-mails worldwide at one point, but specialists say that far more disruptive worms could potentially be let loose.

Worms spread much faster than computer viruses — whereas viruses need to piggyback on other programs in order to propagate, worms simply self-replicate.

Computer departments in most universities and research laboratories had been on the alert since 16 July, when Microsoft announced a flaw in a connection protocol, called the remote procedure call (RPC), used by every Windows machine linked to the Internet. This flaw was quickly exploited by Blaster.

SoBig only spread when users opened the e-mail attachment in which it was hidden. Although the worm saturated networks and slowed down the Internet by creating huge volumes of e-mail traffic, it was widely spotted by antiviral filters and wary users, and only infected about 100,000 machines.

Far more worrying for computer experts is the potential trajectory for 'autonomous network worms' such as Blaster. Instead of



Worm war: recent network attacks have seen users stocking up on computer-protection software.

arriving in e-mails, these worms crawl the Internet, scanning millions of computers for security weaknesses — such as that in the RPC. When they find one, they hack in and replicate themselves. Users are often unaware that their machines have been infected.

In a paper published last year, scientists at the California-based Cooperative Association for Internet Data Analysis (CAIDA) predicted the emergence of high-speed worms that could hijack millions of Internet computers within minutes (S. Staniford, V. Paxson and N. Weaver "How to Own the Internet in Your Spare Time" in *Proc. 11th USENIX Security Symp.* 149–167; USENIX, Berkeley, 2002).

The paper's conclusions, based on mathematical models of existing worms, were partially borne out in January when a worm called Slammer infected almost all 75,000 vulnerable machines within minutes.

Relative to this potential, Blaster was something of a damp squib, says Nicholas Weaver, a researcher at the University of California, Berkeley, and a co-author of the CAIDA study. He points out that the RPC flaw was a "sitting target" for a very large Internet attack by a fast worm, given the huge number of vulnerable machines. Slammer, in contrast,

could only hit the relatively small number of machines hosting Microsoft databases.

"What was remarkable about Blaster was how little damage it did," he says. "The lack of damage was mostly good luck in that Blaster was so poorly engineered." It was "glacially slow", he adds, which gave computer departments time to build defences against it. "A few tricks and it could have spread within minutes," Weaver warns, adding that, like most worms to date, it did not carry a particularly malicious payload. A properly engineered RPC-targeted worm carrying a destructive payload might have blacked-out computer systems worldwide, he claims.

Specialists in assessing that sort of risk will gather in Washington next month for the first Workshop on Rapid Malcode to discuss possible technical responses. But for Bruce Schneier, co-founder of Counterpane Internet Security of Cupertino, California, the problem is not so much technical as legal. He wants software suppliers to be held accountable in court for security problems. "When Firestone produces a tyre with a systemic flaw, they're liable," Schneier says. "When Microsoft produces an operating system with systemic flaws, they're not liable. That's crazy."

Retraction ends furore over cancer vaccine

Alison Abbott, Munich

The authors of a paper, published three-and-a-half years ago in *Nature Medicine*, that claimed that patients with terminal kidney cancer had been successfully treated with individualized cancer vaccines, have finally retracted their work.

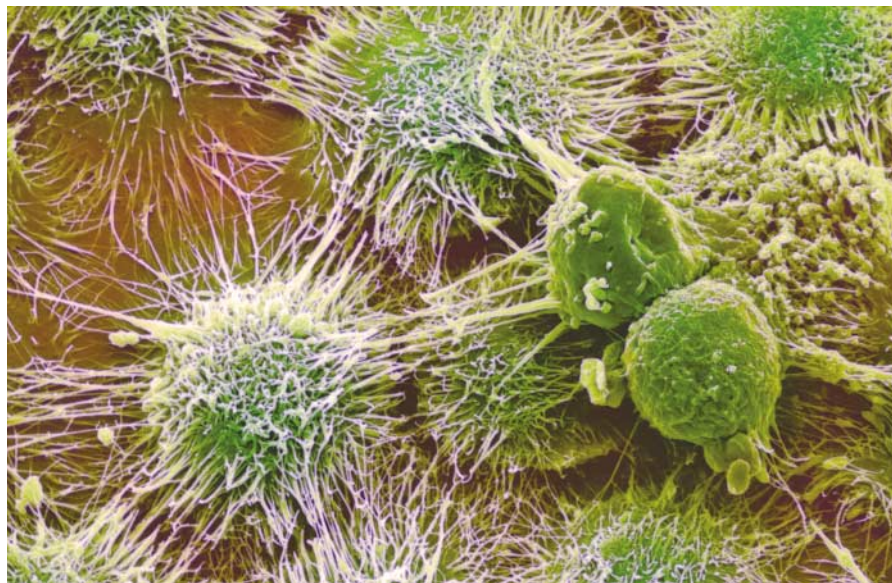
Scientists in Germany, where the clinical studies were carried out, say they are relieved that the long-drawn-out affair is closed. But they claim that the delay in correcting the record has sapped researchers' morale.

An investigation committee at the University of Göttingen reported last November that the paper "failed to meet the requirements of good scientific practice". It found the lead author, Alexander Kugler, guilty of gross negligence, but cleared the other 14 co-authors of scientific misconduct (see *Nature* **420**, 258; 2002).

The contested vaccine trials, sponsored by the Bad Homburg-based healthcare company Fresenius, were suspended more than a year earlier following allegations of irregularities in clinical practice, and the university had been accused of being slow to complete its investigation (see *Nature* **417**, 6; 2002).

The retraction appears in this month's issue of *Nature Medicine*, with an editorial explaining why it took the journal a further ten months from the committee's ruling to persuade the reluctant authors to retract and to agree a form of words (*Nature Med.* **9**, 1221; 2003). "We wanted the authors themselves to retract rather than issuing a retraction from the journal," says editor Beatrice Renault. "It is more meaningful if authors take the responsibility, and it has more power within the scientific community."

One of the authors, Rolf-Hermann Ringert of the University of Göttingen, says that despite errors in the publication —



Cancer cells (green) in kidneys: at the centre of a misconduct case involving vaccine trials.

which include inaccuracies in the primary data and inclusion of patients who did not fulfil the requirements of the trial — he tried for a long time to persuade the journal to publish a list of detailed corrections, rather than a retraction. He says that he still stands by the authors' central claim to have proven the principle that this type of 'fusion' cancer vaccine is effective.

Vaccines of this sort fuse patients' tumour cells with healthy dendritic cells, which present antigens to the immune system in a way that signals that the antigens are foreign and should be destroyed. When the patients are injected with the fused cells, their immune system should be alerted to destroy the cancer cells that they had previously tolerated.

Others working on fusion vaccines say that the Kugler case has had a negative effect

on funding in this area and has made it harder to publish results. But groups in the United States and Japan have now reported anticancer activity of modified fusion vaccines in early-phase trials in breast cancer and in glioma, a type of brain tumour. Fresenius is planning a further trial on kidney cancers at the University of Göttingen, applying optimized fusion technologies, a spokesman for the company says.

Ulf Rapp, a cell biologist at the University of Würzburg and head of a task force that investigated an earlier, unrelated scientific misconduct case, says that the retraction "is the right outcome, but it could have come faster. Journals owe it to their image to take matters into their own hands when an investigation exposes errors and authors are hard to track down, or reluctant to retract." ■

Columbia inquiry prompts White House strategy review

Tony Reichhardt, Washington

The White House will conduct a review of US space strategy in the wake of last week's hard-hitting report on the loss of the space shuttle Columbia — and members of Congress have joined NASA in pledging to reform the space agency itself.

But early soundings in Washington suggest that the manned-spaceflight programme will face an uphill struggle to attract the extra resources that, the report implied, it needs to operate more safely.

The Columbia investigation, led by retired Navy admiral Harold Gehman, was unsparing in its criticism of NASA, as expected (see *Nature* **424**, 863; 2003).

The report also castigated the White House and Congress for underfunding the shuttle programme, which had its budget cut by 40% in real terms during the 1990s.

In a press conference held on 26 August to unveil the report, Gehman called for "a very vigorous public-policy debate about what do we do now" in space, and warned that human spaceflight, if done safely and well, will require much more resources than the United States has been spending. "This stuff is not cheap," he said.

But in the immediate aftermath of the report, few congressional leaders were ready to pledge more money to the space programme. Sherwood Boehlert

(Republican, New York), who chairs the House of Representatives science committee, says that if NASA "is expecting us to write a blank cheque, we're unwilling to do so". Nor is it certain, he adds, that the astronaut programme will continue. "My own predisposition is to continue manned spaceflight, but not at any cost, and not at any risk." Any push to increase NASA funding will have to come from the Bush administration. "We're going to have to have clear direction from the White House," says Boehlert.

Last year the White House Office of Science and Technology Policy and the National Security Council led a review of US

Peers rally in support of accused scientist

Erika Check, Washington

Two leading US scientific organizations have lambasted the government's treatment of an infectious-disease expert who faces multiple criminal charges after mishandling plague samples.

In a letter to Attorney General John Ashcroft, Bruce Alberts, the president of the National Academy of Sciences, and Harvey Fineberg, who heads the Institute of Medicine, said that the prosecution of the researcher, Thomas Butler of Texas Tech University in Lubbock, was endangering US bioterrorism research.

Butler was arrested in January for giving incorrect information about samples of the plague bacterium *Yersinia pestis* he had transported to Texas from Tanzania. Initially, Butler, who was chief of infectious diseases at Texas Tech, had reported that the samples were lost, and then confessed that he had destroyed them. The US Department of Justice charged Butler with illegally transporting the vials, lying to federal agents, and filing a false 2001 tax return.

If convicted on all current counts, Butler could pay as much as \$3.6 million in fines and go to prison for 74 years, according to the National Academies' Committee on Human Rights. Additional charges are expected to be filed before his trial, set for 6 October in Lubbock.

In their letter to Ashcroft, Alberts and Fineberg wrote: "We are particularly concerned about the impact that Dr Butler's case may have on other scientists who may be discouraged from embarking upon or continuing crucial bioterrorism-related scientific research — thereby adversely affecting the nation's ability to fully utilize such research capabilities in preparing

space policy that was nearly complete at the time of the Columbia accident. That project was shelved but now is being revived with the Gehman report in mind, for completion early next year, just in time to feed into the Bush administration's 2005 budget request.

Meanwhile, NASA may report as early as this week on its latest plan for returning the shuttle to flight. The earlier target launch date of next spring now looks unlikely, as does the pre-accident plan to complete the US segments of the International Space Station within a year of the next shuttle launch. The Gehman panel's report cited pressure to finish the station on schedule as a contributing cause of the accident.



Plagued: scientists believe Thomas Butler's treatment will deter others from biodefence research.



defenses against possible bioterrorist attacks."

On 28 August, the human-rights committee wrote to its membership — 1,700 senior scientists, engineers and doctors — asking them to support Butler by writing to Texas Tech, the US government and Butler. It also requested donations for Butler's defence

The current US federal deficit and the mounting costs of the war in Iraq make any grand expansion of NASA's budget unlikely, but some commentators say that political factors may be coming together for a renewed US commitment to human space exploration. Howard McCurdy, a space-policy expert at the American University in Washington, says that President Bush may be under pressure to make such a commitment. "Historically," he says, "no major political figure with the authority to make the decision has ever wanted to put his or her fingerprints on a decision to start closing down the manned-spaceflight programme."

— so far, he has paid his lawyers \$400,000, the alert says.

Nobel laureate Torsten Wiesel, who chairs the human-rights committee, says that the only other case in which the committee has taken similar steps on behalf of a US scientist was the prosecution of Wen Ho Lee, the Los Alamos National Laboratory researcher who was accused of stealing US nuclear secrets in 1999.

But the committee was compelled to act in Butler's case because of what seemed to be a severity of punishment out of proportion to his crime, Wiesel says. "Rather drastic steps have been taken against him, and we want to ensure that he is being dealt with properly by the justice department and the university," he explains.

Texas Tech has prevented Butler from returning to its campus at Lubbock and is planning to testify against him at his trial; Butler, meanwhile, is becoming depressed, Wiesel says.

Wiesel also warns of the case's broader significance. He says that several scientists have told the National Academies that the Butler prosecution is scaring US researchers away from biodefence work, because they perceive Butler to be a scientist who has merely been caught unaware by the rules.

"It's a serious mistake because he is one of the internationally recognized experts in plague," Wiesel says. "This has certainly sensitized the community, which is not helpful if you really are concerned about bioterrorism."

Steven Block, a biophysicist at Stanford University, adds: "The Department of Justice should go after individuals who pose serious, credible threats to our national security, not hapless biomedical scientists who are simply pursuing their research, but failing to fill out the mounds of paperwork mandated by the bureaucracy."

Court ruling sounds note of caution for sonar system

Rex Dalton, San Diego

A court has ordered the US Navy to come up with an environmental plan before it extends the deployment of a submarine-detection system that some biologists say could disturb and harm marine mammals and fish.

The ruling by a federal judge in San Francisco could end a prolonged wrangle between the navy and environmentalists over the use of the sonar system (see *Nature* 413, 243; 2001).

Environmental groups and the navy will meet on 7 October to work out terms for deploying the new sonar, which emits a low-frequency, 140-decibel sound wave that is able to pick up echoes from submarines hundreds of kilometres away.

On 26 August, Judge Elizabeth Laporte of the US District Court in San Francisco ruled that the navy did not have proper environmental permits to test and use the sonar system widely. Environmentalists greeted the ruling as a major victory in their long-running efforts to block its deployment. The navy, even as it complies with the ruling, will push Congress to change the law.

In her 73-page ruling, Laporte said that the case required a difficult balancing act between national security requirements and laws enacted to protect sea life. She declined to ban the sonar outright, but sought a permanent, court-approved plan to allow broader use of the equipment while protecting sea life.

The decision "recognizes that during peacetime even the military must comply with our environmental laws", says Joel Reynolds, a senior attorney for the Washington-based National Resources Defense Council, the environmental group that led the lawsuit. The navy said that it was "concerned about the implications of the decision for national defence".

Currently, the navy can only test and train with the new sonar system in a region covering two million square kilometres near Guam in the western Pacific Ocean. But the navy wants to use it in ocean habitats that are populated by whales and other species that are already suffering environmental pressures.

Other sonars that operate at higher frequencies have been linked to incidents in which whales, dolphins or porpoises were harmed or killed (see *Nature* 415, 106; 2002).



Relatives gathered in Ayacucho last week to hear Truth Commission findings on Peru's years of conflict.

Statistical model leaves Peru counting the cost of civil war

Jonathan Knight, San Francisco

A statistical method for estimating animal populations is proving increasingly useful in documenting human-rights abuses. Its latest achievement, revealed last week, was to reconcile conflicting accounts of the death toll from two decades of civil war in Peru.

Up to 35,000 people were believed to have been killed in a conflict that raged between the government and Shining Path Maoist insurgents from 1980 until about 2000. The new analysis, part of the final report from Peru's Truth and Reconciliation Commission delivered in Lima on 28 August, concludes that more than 69,000 people died. Of these, 46% were killed by the Shining Path, 30% by the government, and the rest by other rebels and paramilitary groups.

"As soon as people hear these numbers, they ask how could it have been so many, and why did they have no idea," says sociologist Patrick Ball, the study's lead author and deputy director of the Science and Human Rights Program at the American Association for the Advancement of Science (AAAS).

Accurate death counts during civil conflicts are hard to come by. Bodies may be disposed of anonymously in mass graves, and witnesses may distrust investigators. In Peru, much of the killing took place in remote villages in the Andes.

Ball and his colleagues applied a statistical method developed in the late nineteenth century to count wild animals. It has since been used to adjust census estimates of hard-to-count groups such as the homeless.

Multiple-systems estimation (MSE) relies on the existence of overlapping partial

counts of a population. The principle is that the chance of an individual appearing on two lists is equal to the product of their chances of appearing on each list separately. Ball's team compared seven lists of dead and missing people compiled since 1980.

The researchers had to factor in potential sources of bias. For example, one public agency was charged with collecting testimony about abuses by the military, so its databases contained little about murders by rebels. Yet its statistics had been touted as evidence that the government did most of the killing. "People weren't thinking like statisticians," says Ball. "They were throwing the numbers around without thinking about the process."

This is not the first time MSE has proved its worth. In 1999, Ball wrote a AAAS report on the 1981–83 conflict in Guatemala which revealed a disproportionate number of deaths among people of Mayan descent, ultimately leading a United Nations commission to conclude that the Guatemalan army had engaged in genocide. Last year Ball testified at the trial of former Yugoslav president Slobodan Milosevic that the same method confirmed the Yugoslav army as responsible for the deaths of thousands of Albanians in Kosovo in 1999.

Although many leaders of the Shining Path are already dead or in jail, the military has never been held accountable, says Sergio Meza, head of the Peruvian section of Amnesty International: "We now have the opportunity to prosecute those members of the military who violated human rights." But there is strong political and military opposition to such prosecutions.

Lion man takes pride of place as oldest statue

Rex Dalton, Reno

Intricate ivory carvings said to be the oldest known examples of figurative art have been uncovered in a cave in southwestern Germany. Researchers say that the finding could change our understanding of early man's imaginative endeavours.

The artefacts — including a figurine depicting a *Lowenmensch* ('lion man') — have been carbon-dated to around 30,000 years ago, when some of the earliest known relatives of modern humans populated Europe.

Discovered last year by a team led by US archaeologist Nicholas Conard of the University of Tübingen in Germany, at the Hohle Fels cave near Ulm, the objects include figures depicting a horse and a bird.

Conard says he thinks that the figures are older than a previously discovered *Lowenmensch*, fragments of which were found by German archaeologists in 1939 near Vogelherd and dated to about the same time. Until now, those artefacts were accepted as the oldest examples of figurative art in the world. The newly discovered objects are older, Conard argues, as they were uncovered at a lower level in the cave floor's sediments.

"These discoveries have incredible significance," says Clive Gamble, an archaeologist at the University of Southampton, UK. "They depict the animal world in a semi-realistic way. It shows early man moving from his immediate world to an imaginative world."



The Vogelherd 'lion man' could lose its crown to another as the oldest figurative work of art.

Conard reported the discovery on 30 July at the Sixteenth Congress of the International Quaternary Association in Reno, Nevada, during a lecture on the human colonization of Europe. He is now preparing a manuscript on the discovery for publication.

Conard, who has studied human migra-

tion from Africa at dig sites stretching from Syria to Germany, believes that humans first arrived in central Europe by following the River Danube west into the area. The figurines add a new dimension to theories about the Danube route, agrees Gamble. "During the Ice Age in Europe, the frozen Danube would be like a highway," he says, "providing a fast track to new environments."

Fossil remains suggest that modern humans and Neanderthals both lived in Europe during this period. Conard reported that the sedimentary levels in which the ivory carvings were embedded did not include any Neanderthal fossils. But some archaeologists argue that it is possible that the much-maligned Neanderthals produced similar objects. "I don't think that is as far-fetched as some people might think," says Jeffrey Brantingham, an archaeologist at the University of California, Los Angeles. "These objects are pushing the markers and traits" of modern man "further back into time", he says.

Archaeologists have pointed out that beads, bone points and pendants have already been discovered in association with Neanderthal fossils. Attributing artefacts to one of the two hominid groups remains difficult, says Brantingham. Gamble says that the discovery will spur fresh exploration of France, Spain and South Africa, where even older cave drawings — but not figurative art of this age — have been identified. ■

Britain set to pull plug on nuclear-fuel reprocessing

Quirin Schiermeier

The United Kingdom looks set to get out of the nuclear-fuel reprocessing business, after a government discussion paper suggested that British Nuclear Fuels (BNFL) shift its emphasis to nuclear decommissioning and waste disposal.

The suggestion has prompted a wave of speculation that BNFL, which runs the Thorp (thermal oxide reprocessing) plant at Sellafield, Cumbria, will close it by 2010, when its current contracts will have expired.

BNFL officials say that no final decision had been made to close the Thorp plant, which cost £1.8 billion (US\$2.8 billion) to build and only opened in 1994. But David Bosner, the company's deputy chief executive, acknowledges that Thorp has no more orders after 2010. "Beyond the point that current customers have placed orders, every industry has doubts about what its future's going to be, and we're no different," Bosner told the BBC.

BNFL is owned by the UK government, but will be taken over by a new Nuclear

Decommissioning Authority, responsible for cleaning up Britain's nuclear sites, in 2005.

Spent fuel from nuclear power stations is separated at Thorp into uranium, plutonium and highly radioactive fission products. Uranium and plutonium are then recycled for possible re-use as nuclear fuel. The remaining high-level nuclear waste is turned into a radioactive glass and placed in metal canisters for storage.

Thorp will lose much of its workload when a Japanese reprocessing plant opens for business in 2005. If Thorp closes, Japan and France will be the only countries left pursuing a complete nuclear fuel cycle of the sort widely envisaged when civilian nuclear power emerged 50 years ago. The United States abandoned the fuel cycle in the 1970s following criticism that it produced materials suitable for use in nuclear weapons.

Britain had seen nuclear reprocessing as a money-earner, but business has failed to materialize, and Thorp operates at just 50% of its capacity. "Reprocessing is a profitable business as long as uranium prices are high,"



The end could be in sight for Britain's nuclear-fuel reprocessing plant at Sellafield.

says Ian Hore-Lacy, a spokesman for the World Nuclear Association, a London-based lobby group. "But prices have been low for over a decade because of uranium supplies from Russian weapon stockpiles."

In addition, several European countries are about to ban the export of spent nuclear fuel for reprocessing, points out Heinz Sager, a spokesman for NAGRA, the Swiss National Cooperative for the Disposal of Radioactive Waste. ■

Biotech headquarters hit by early morning pipe-bomb explosions

San Francisco Animal-rights activists have claimed responsibility for two pipe bombs that exploded outside a Californian biotechnology company last week.

The blasts on 28 August occurred about an hour apart before dawn at the Emeryville headquarters of the Chiron Corporation. Windows were broken but no one was injured. Most of the 2,000 employees were asked to stay at home while investigators combed the scene.

A group calling itself Revolutionary Cells — Animal Liberation Brigade said in an e-mail to local media that it had planted the bombs to protest against Chiron's use of animals in drug testing. Previous animal-rights criminal activity in the United States has mainly involved arson and vandalism.

Conservation plans stymied by inadequate data

London Good intentions about biodiversity conservation are of little use without data to back them up, the Royal Society has warned.

National governments committed themselves last summer to achieving a

significant reduction in the rate of loss of biodiversity by the end of the decade. But this target, agreed at the World Summit on Sustainable Development in Johannesburg, cannot be met without the development of better systems for assessing numbers of plant and animal species, according to a report released by the society this week.

The authors describe a framework that could be used to unify such assessments, from the identification of who is interested in the biodiversity of an area, to pilot studies for testing data-gathering strategies.

"We need to be able to tell if the loss of species and their habitats is speeding up or slowing down, and whether conservation is having any impact," says Georgina Mace, director of science at the Zoological Society of London and an author of the report.

Fungus work flourishes in ashes of forest fires

Munich This summer's forest fires in Europe have let scientists gather fresh data on one of biology's most important model organisms — the fungus *Neurospora crassa*.

The fungus has an 80-year track record as a model organism, and its full genome sequence was published in April (J. E. Galagan *et al. Nature* **422**, 859–868; 2003). It is already used in genetics, photobiology



Silver lining? The fires that devastated Europe may open up uncharted areas of biology.

and circadian biology, but *Neurospora* fans say that its use could be extended to new areas of biology, including ecology and evolution, if more were known about its natural history.

After the fires, researchers from the University of Munich began trawling European forests for new samples of the fungus, taking advantage of the fact that it rapidly colonizes vegetation that has been killed in a blaze. Similar initiatives have been run in other countries during the past few years, and samples are being sent to Stanford University in California for analysis.

"The fires here were appalling," says Martha Merrow, a geneticist at the University of Munich who is championing the European initiative. "But a small good may come out of it for science."

Biology favoured in Japanese funding plan

Tokyo The life sciences look set to prosper while engineering research is squeezed, under budget proposals released by Japan's science ministry last week.

Japan's budget process involves ministries pitching high, projecting what they would like to spend. The government — guided, in the case of science spending, by its Council for Science and Technology Policy — will pare down these proposals to match the limited funds available in 2004.

In its proposal, the Ministry of Education, Culture, Sports, Science and Technology seeks an increase of 21%, taking its research spending to ¥949 billion (US\$8 billion). It won't get that much, observers say, but its proposal shows where priorities lie: life sciences gets 26% and nanotechnology 35%, for example, whereas nuclear research gets 10% and aerospace programmes just 9%.

Michigan unveils major life-sciences institute

Washington Michigan's bid for prominence in biology research will get a shot in the arm when a \$230-million research facility opens next week.

The goal of the Life Sciences Institute, at



Funding for carbon nanotube research could increase under Japanese budget plans.

the University of Michigan in Ann Arbor, is to gather biologists, physicists and chemists together to foster innovative approaches to biological problems, says spokesman Karl Bates. Similar approaches are being taken at Stanford University's Bio-X project and Harvard University's Bauer Center.

The institute is good news for Michigan's efforts to become a biotechnology powerhouse, which have been scaled back because of the state's budget problems (see *Nature* **422**, 102; 2003). The new institute has a dedicated endowment of \$130 million, unaffected by the cutbacks.

► www.lifesciences.umich.edu

Universities armed with cash for biotechnology

Washington The US Army is to give three universities \$50 million over five years to develop military products using biotechnology.

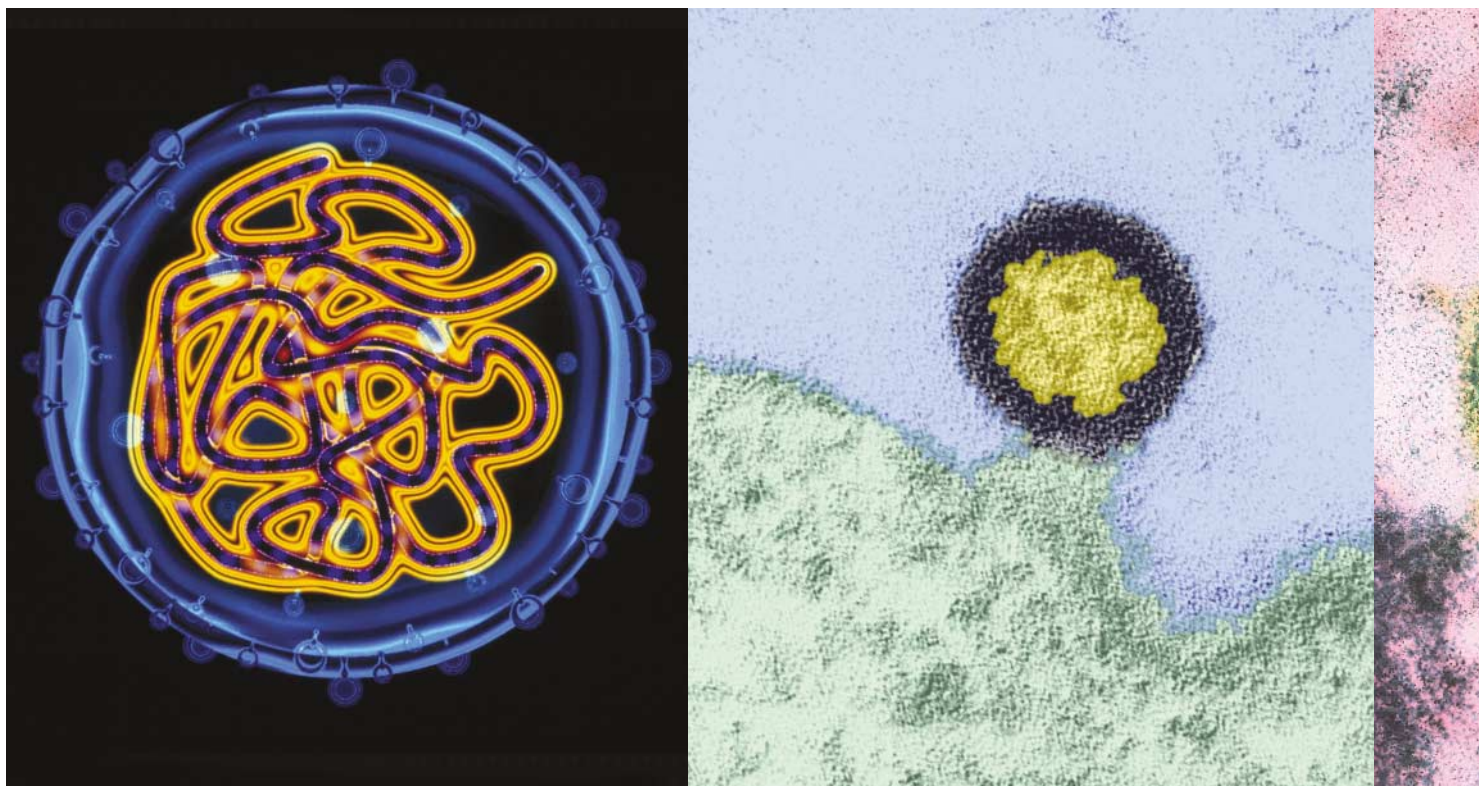
The Institute for Collaborative Biotechnologies, to be led by the University of California, Santa Barbara (UCSB), will focus on three areas: sensors and electronics, biologically inspired new materials, and biological methods of information processing.

"The inspiration for the institute comes from the fact that biology uses precise mechanisms to produce exquisitely structured materials," says Robert Campbell of the Army Research Office in Research Triangle Park, North Carolina.

UCSB will work with the Massachusetts Institute of Technology and the California Institute of Technology on the project.

Correction

An article about research funding at the newly formed US Department of Homeland Security incorrectly stated that the agency had received over 3,000 unsolicited bids for research projects since it opened in January (see *Nature* **424**, 986; 2003). In fact, that is the number of solicited proposals the agency received in its first call for grant applications, which closed on 13 June.



RNA to the rescue?

Disease therapies based on a technique for gene silencing called RNA interference are racing towards the clinic. Erika Check investigates molecular medicine's next big thing.

Medicine's molecular revolution is overdue. By now, enthusiasts led us to believe, gene therapy and related treatments should have transformed clinical practice. Diseases, they told us, would be cured at their genetic roots, by repairing defective human DNA or by disabling the genes of infectious microbes. But it has proved frustratingly difficult to make these methods work in the clinic — if you get sick, your doctor will probably still treat you with the pills and potions of old-fashioned medicinal chemistry.

Given this chastening experience, you would expect experts to be cautious about the prospects of molecular medicine's latest hope — a gene-silencing mechanism known as RNA interference, or RNAi. But instead, researchers can barely contain their enthusiasm. "Right now, everybody's excited," says Anastasia Khvorova, director of biology with Dharmacon in Lafayette, Colorado, a company that supplies RNAi technologies to researchers and companies that develop therapeutics.

The term RNAi was coined just five years ago, in a paper documenting the phenomenon in the nematode worm *Caenorhabditis*

*elegans*¹. Yet doctors and biotech executives are now talking about beginning human trials within the next two or three years — an astonishing rate of progress. Part of the excitement stems from the knowledge that, unlike techniques such as gene therapy, RNAi is a natural defence mechanism that is thought to have evolved to protect organisms from viral diseases.

Dicey defence

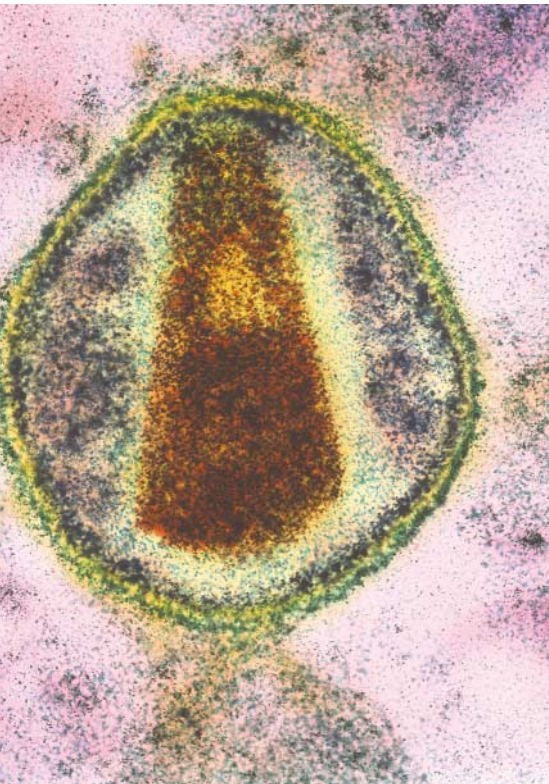
Many viruses have a genetic blueprint made from RNA, rather than DNA. When they infect a cell, they make double-stranded copies of their genetic material. In response, the RNAi pathway strikes back. An enzyme known as Dicer first chops the double-stranded viral RNA into small segments of genetic code, each around 22 'letters' long. These segments, known as small interfering RNAs, or siRNAs, then separate into single strands and some bind to intact stretches of single-stranded viral RNA. Finally, proteins target this tagged viral RNA and destroy it². As a result, RNAi shuts off key viral genes, potentially nipping infections in the bud.

Biologists are exploiting RNAi as an experimental tool to find out what genes do.

When a gene is activated, its sequence is read to produce messenger RNA (mRNA), which contains the information necessary to manufacture a particular protein. So by using siRNAs or double-stranded RNAs that correspond to a specific mRNA sequence, researchers can trick a cell into destroying this mRNA and silencing the gene in question.

As soon as it became obvious that the phenomenon operates in mammals³ as well as in lower organisms, clinicians pricked up their ears. In theory, RNAi could be used to treat any disease — forms of cancer, for instance — that is linked to an overactive gene or genes. But for the time being, most of the clinical interest lies in applying RNAi in its natural role: as a means of combating pathogenic viruses by disabling their RNA.

One of the obvious targets is HIV — a virus for which there is no cure and no vaccine. Last year, for instance, molecular virologist Bryan Cullen of Duke University Medical Center in Durham, North Carolina, introduced siRNAs against two HIV genes into the human immune cells that are destroyed by the virus. The siRNAs allowed these cells to resist viral replication better than those that had not been triggered to undergo RNAi



Combating the incurable: researchers are testing the idea that the RNAi pathway, which shuts down the genes of invading viruses, can block the replication of hepatitis C virus (far left, RNA shown in yellow) and HIV (left and middle left).

In June, this company merged with Ribopharma of Kulmbach in Germany.

Both Avocel and Alnylam are planning to begin clinical trials as early as 2005. And hot on their heels is Sirna Therapeutics of Boulder, Colorado, which has already raised US\$43 million from investors. Sirna aims to develop therapies for hepatitis C and an eye condition called macular degeneration. For now, these companies are maintaining amicable relations. But the situation could get messier as RNAi moves towards the clinic, because patent offices around the world have not yet decided who owns the rights to some key RNAi-based technologies.

Before worrying about the ownership of key intellectual property, however, scientists must figure out how to make RNAi therapies work. They are facing some formidable technical barriers, chief among which is the problem of getting siRNAs into the right cells. This is not a trivial issue, because RNA is rapidly broken down in the bloodstream, and our cells don't readily absorb it through their membranes. And even when RNA gets into its target cell, scavenger proteins quickly chew it up. "The major hurdle right now is delivery, delivery, delivery," says Sharp.

Researchers are exploring a variety of ways to combat the problem. Some involve techniques developed to facilitate an older technology known as 'antisense'. The idea behind antisense is to muffle a cell's single-stranded mRNA — the 'sense' strand — using a piece of antisense RNA with a 'complementary'

sequence that binds tightly to the mRNA. This, the theory goes, should prevent the mRNA from being translated into protein. Scientists tried a variety of ways to get the antisense RNAs into cells — for example, they packed the RNA inside fatty globules, called liposomes, which can cross cell membranes. But antisense has not performed well in clinical trials, partly because these delivery systems were not particularly effective. Khvorova believes that the medical benefits of RNAi will be huge if the delivery issues can be resolved. "But we've looked at a lot of the delivery methods that have been used for antisense, and so far I haven't been impressed," she says.

Harmless HIV

Another option is to use a harmless virus as a vector to ferry RNAi-triggering genes into their target cells. Molecular biologist John Rossi of the Beckman Research Institute of the City of Hope Medical Center in Duarte, California, is experimenting with one such vector, based on a version of HIV from which the disease-causing genes have been stripped. Together with colleagues led by Ramesh Akkina of Colorado State University in Fort Collins, Rossi engineered this vector to contain sequences encoding siRNAs targeted against HIV genes. The researchers used their vector to infect the human stem cells that develop into immune cells. Next, they either grew the cells into mature cells in the lab, or injected them into mice from a special strain that accepts human transplants. In both cases, the mature immune cells fought off HIV when researchers tried to infect them with disease-causing HIV in culture dishes¹⁰.

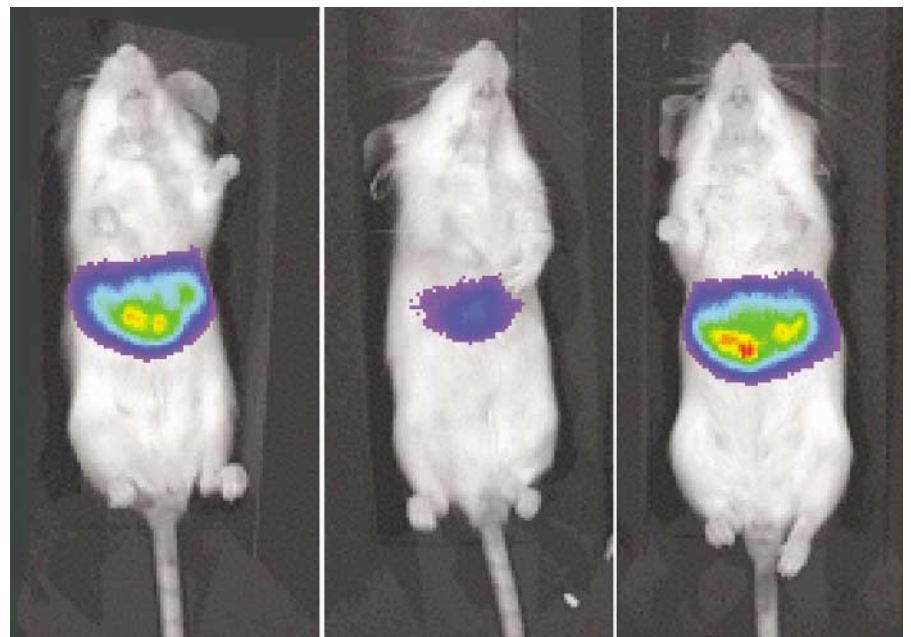
Rossi hopes that a similar technique could work in human patients with HIV. Doctors

(ref. 4). Meanwhile, other researchers have shown that, in cultures of human cells, RNAi can similarly combat viruses as diverse as respiratory syncytial virus⁵, and those that cause influenza⁶ and polio⁷.

RNAi may work like a charm in petri dishes — but what about in live animals? Mark Kay, a geneticist at Stanford University in California, addressed this question by fusing a genetic sequence from the hepatitis C virus to a gene for the enzyme luciferase, which stimulates a reaction that emits light. When Kay injected the fused gene into mouse livers, he could track its location by detecting the glow. And when the mice were treated with siRNAs targeted against the hepatitis C gene, this glow dimmed dramatically⁸. Hepatitis C doesn't make mice sick, but Kay and his colleagues have since gone on to show that RNAi can drastically reduce signs of infection by hepatitis B (ref. 9), which can damage the animals' livers.

Firm plans

Results such as these are attracting intense commercial interest. In August, Kay announced that he has licensed his work on hepatitis C to a company called Avocel in Sunnyvale, California, which aims to develop RNAi therapies against the disease. Other RNA pioneers are lining up with their own start-up biotech firms. For instance, Phillip Sharp of the Massachusetts Institute of Technology in Cambridge, who shared the Nobel Prize in Physiology or Medicine for his earlier work on RNA splicing, is one of the co-founders of Alnylam Pharmaceuticals, also based in Cambridge.



In mice containing a glowing version of a hepatitis C gene (left), a small interfering RNA (siRNA) against the gene reduces liver fluorescence (middle), but an unrelated copy of the siRNA (right) does not.

REF. 8

would extract stem cells from a patient's bone marrow, infect them with the RNAi-triggering vector, and then put them back into the patient. Rossi is now working to perfect this technique in mice, and is also beginning tests in rhesus monkeys to ensure that the treatment has no unwanted side effects. He hopes to convince the US Food and Drug Administration to authorize a clinical trial in the next two or three years. "I think when we get all this data compiled we'll have a fairly free road into a stem-cell trial," Rossi predicts.

This may be so, but there are nagging safety concerns about vectors made from viruses in the same family as HIV, which are called retroviruses. This is due to the fact that retroviruses work by forcing their way into a cell's own DNA. If the vector lands in the wrong place it can damage important genes and even cause cancer. These concerns were borne out by last year's revelation that a retroviral vector had triggered leukaemia in some children in a gene-therapy trial¹¹. Because of these concerns, Rossi says that he will not use stem cells in his first clinical trial. Instead, he will initially treat mature immune cells, because these cells are less likely to grow out of control.

Safe delivery

Kay, meanwhile, is pinning his hopes for an RNAi vector on a virus known as adeno-associated virus, or AAV. He has already used AAV-based vectors in clinical trials of gene therapy against haemophilia¹². AAV does not cause disease in people, and so far there has been no cause for any serious safety concern — even though AAV can also integrate into a cell's own DNA.

Another important question mark hanging over RNAi is its specificity. Before regulators give the go-ahead for a clinical trial, scientists need to prove that that RNAi will not shut down vital human genes as well as the target viral sequences.

Some studies on specificity have yielded encouraging results. In May this year, for instance, researchers led by Patrick Brown of Stanford University reported on experiments in which they engineered human kidney cells to produce a fluorescent protein. They shut down the gene for this glowing protein by using RNAi, and then used DNA microarrays to monitor some 20,000 other genes — none of which seemed to be affected by the treatment¹³.

But just a couple of weeks later, researchers with Rosetta Inpharmatics in Kirkland, Washington, cast a shadow over this rosy picture. The Rosetta team, led by Aimee Jackson and Steven Bartz, used a range of different siRNAs to target two genes in cultured human cells. Disturbingly, the treatment caused changes in the expression of dozens of other genes. Depending on the precise sequence of the siRNA concerned, a different range of 'off-target' genes seemed to be affected¹⁴.



Little helpers: Mark Kay hopes to use harmless viruses to deliver RNAi therapy to patients.

Jackson and Bartz are not sure why their results were so different from those obtained by Brown's team, but one possible explanation is that they used larger doses of siRNA. The Rosetta researchers also tested for off-target effects sooner after beginning their experiment than other groups have in their studies. But whatever the explanation, the findings have shaken up the RNAi camp. "We've had some really lively discussions," says Bartz.

New data from a group at Case Western Reserve University in Cleveland, Ohio, seem to support the Rosetta findings¹⁵. Last week, Bryan Williams and his colleagues reported that when they introduced siRNAs into cells, certain genes that are part of the interferon

system were activated, a mechanism by which cells shut themselves down in response to invading germs. The siRNAs activated genes that act early in the interferon pathway, and Williams' group did not measure whether the activated genes stopped working. But the team says that its findings provide a warning that off-target effects are perhaps more common than scientists have realized.

Researchers argue that these hints of off-target RNAi effects highlight the need for a deeper understanding of how, exactly, the system works. For instance, we still don't know for sure how many proteins work together to shut down a target mRNA. It's also unclear why some siRNAs are incredibly effective, whereas others, targeted at a different region of the same gene, don't work as well. Given these unknowns, some researchers urge caution before rushing into clinical trials. "Before you know what you could perturb, you have to know what's there," says Tom Tuschl, a biochemist and RNAi pioneer at Rockefeller University in New York.

Even some of the scientists working in the commercial sector, where excitement about the clinical prospects of RNAi is most intense, agree that a great deal of ground-work remains to be done. "For real clinical development, this has to be done right," says Khvorova. "Investing a little more time on the basic steps will pay back in years of time saved later on."

But despite all of these caveats, most researchers working in this fast-moving field have high hopes that RNAi will deliver on its therapeutic promise. "This is the honeymoon period; things are looking great," says Kay. "We will encounter technological issues along the way, but our goal is to solve these problems and get it to work."

Erika Check is Nature's Washington biomedical correspondent.



Going to market: Phillip Sharp is one of several RNAi researchers to form start-up biotech firms.

1. Fire, A. *et al.* *Nature* **391**, 806–811 (1998).
2. Hannon, G. J. *Nature* **418**, 244–251 (2002).
3. Elbashir, S. M. *et al.* *Nature* **411**, 494–498 (2001).
4. Coburn, G. A. & Cullen, B. R. *J. Virol.* **76**, 9225–9231 (2002).
5. Bitko, V. & Barik, S. *BMC Microbiol.* **1**, 34 (2001).
6. Ge, Q. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 2718–2723 (2003).
7. Gitlin, L., Karelsky, S. & Andino, R. *Nature* **418**, 430–434 (2002).
8. McCaffrey, A. P. *et al.* *Nature* **418**, 38–39 (2002).
9. McCaffrey, A. P. *et al.* *Nature Biotechnol.* **21**, 639–644 (2003).
10. Banerjee, A. *et al.* *Mol. Ther.* **8**, 62–71 (2003).
11. Check, E. *Nature* **420**, 116–118 (2002).
12. Manno, C. S. *et al.* *Blood* **101**, 2963–2972 (2003).
13. Chi, J.-T. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 6343–6346 (2003).
14. Jackson, A. L. *et al.* *Nature Biotechnol.* **21**, 635–637 (2003).
15. Sledz, C. A., Holko, M., de Veer, M. J., Silverman, R. H. & Williams, B. R. G. *Nature Cell Biol.* doi:10.1038/ncb1038 (2003).

Dharmacon

♦ www.dharmacon.com

Avocel

♦ www.avocel.com

Alnylam Pharmaceuticals

♦ www.alnylam.com

Sirna Therapeutics

♦ www.sirna.com

Nature RNA Insight

♦ www.nature.com/nature/insights/6894.html

Hot tempers, hard core

Were the last dinosaurs wiped out when interplanetary debris hit what is now the coast of Mexico? A drilling project that promised answers has been plagued by squabbles over access to samples. Rex Dalton reports.

It should have been a geological fiesta. In April 2002, an élite group of scientists converged on Mexico City to examine one of the most important cores ever drilled from the Earth.

Since the previous December, local engineers had drilled 1,511 metres down into the Chicxulub crater on Mexico's Yucatán Peninsula, to extract cores of rock. Researchers hoped that these samples would help to solve one of the biggest mysteries in our planet's history — what caused the extinction of the last of the dinosaurs, together with most other species represented in the fossil record, some 65 million years ago.

The now-hidden crater, detected by anomalies in measurements of Earth's gravity and magnetism, was formed at about the same time, when a meteorite or comet hit the planet. Many researchers believe that this impact caused the mass extinction that saw off the dinosaurs by kicking up debris into the stratosphere, blocking sunlight and shutting down most photosynthesis on the planet. A minority disputes this theory, arguing that other events — such as volcanic eruptions, sea-level changes or a series of impacts — were to blame for the spectacular loss of species that occurred at the transition between the Cretaceous and Tertiary periods, a time otherwise known as the K/T boundary.

Samples from the crater should untangle the sequence of events associated with the Chicxulub impact, and help scientists to assess whether it could have caused the K/T extinction. But last April's gathering, at the National Autonomous University of Mexico (UNAM), kicked off a prolonged period of acrimony, as researchers battled for access to segments of the core. In particular, one prominent proponent of the impact theory stands accused of delaying the acquisition of key core samples by other project scientists. And the row has been heightened by the claim from one of those scientists that these disputed samples cast severe doubt on the idea that the Chicxulub impact caused the mass extinction.

"Nothing is more political in science than



High impact: extracted from a depth of 1,500 metres, core samples (left) from the Chicxulub crater offer insight into what caused the dinosaurs to die out.

the Chicxulub project, and the politics is holding back the science,"

observes Doreen Ames, a geologist with Natural Resources Canada in Ottawa, whose work was caught up in the row.

Efforts to drill into the Chicxulub crater suffered a decade of false starts because of funding difficulties. But eventually, the International Continental Scientific Drilling Program (ICDP), based in Potsdam, Germany, agreed to pump US\$1.5 million into a

coring operation. Mexico provided another \$500,000 for the drilling, plus \$200,000 for membership of the ICDP; government agencies in various countries chipped in with cash for research associated with the core.

Drilling began on 9 December 2001, at a site called Yaxcopoil, close to the town of Mérida and some 60 km from the centre of the crater. Despite minor setbacks — including a brief halt because of a financial dispute with the drilling company — by April last year the core had been boxed, shipped to UNAM in Mexico City, and prepared for the arrival of about two dozen of the project's scientists. The core, about 8 centimetres in diameter, was cut into sections a few metres long, and then split lengthwise. Half of each section was archived, and when the participating scientists arrived at UNAM, the plan

ICDP/GEZ POTSDAM

ICDP/GEZ POTSDAM

was to divide the other halves into smaller samples for analysis of their chemical composition, mineral content and microfossils.

All eyes were on the boundary-core section from nearly 800 m deep, split by a thin green line. These deposits of a mineral called glauconite, the researchers knew, were laid down in marine sediments during the K/T boundary. "They looked at that core and started drooling," says Buck Sharpton of the University of Alaska at Fairbanks, one of the Chicxulub project's six principal investigators. With other sections of the core, researchers were able to select their samples, secure them and leave. But with more scientists wanting samples from the boundary section than could be accommodated, things became heated.

So much so, in fact, that two principal investigators — Jan Smit, a geologist at the Free University in Amsterdam, and UNAM team leader Jaime Urrutia Fucugauchi — clashed over Smit's plan to take the boundary section of the core back to the Netherlands and distribute it to project scientists from there. Smit argued that he could do a better job of cutting the core, which irritated the Mexican scientists. "I was very angry," says UNAM geologist Mario Rebolledo Vieyra.

Rock stars

Eventually, Urrutia relented. In June 2002, Smit took the boundary-core section to Amsterdam to cut specimens for analysis. The agreement, confirmed by Smit in a 3 July letter to Urrutia, was that Smit would distribute boundary-core samples to project scientists worldwide by 15 August, and return the remaining core material to Mexico by 15 September. But those deadlines were missed, and Urrutia says that he regrets entering into the agreement. "I now believe Smit wanted to control that segment," he claims.

Smit categorically denies this allegation and blames a busy schedule and poor communication from the scientists wanting K/T boundary samples. Many of them, as it happens, were e-mailing or calling Mexico, believing the core section to be there. Urrutia admits that his team should have done a better job of informing project scientists that the samples were in Smit's lab. "There was a lot of misunderstanding," he says. "In retrospect, we should have communicated better." For his part, Smit concedes that he "waited too long" to send out the boundary-core material — one-third of the core, which should have been returned to UNAM, was still in his lab as *Nature* went to press.

As the months of 2002 slipped by, about a dozen researchers waiting for their boundary-core samples became progressively more nervous. They had another deadline to worry about — the 15 January 2003 cut-off for the submission of abstracts for a symposium on the Chicxulub project, scheduled for a joint meeting of the American Geophysical Union,



Cutting edge: the Chicxulub drill site is located near Mérida in Mexico's Yucatán Peninsula.

the European Geophysical Society and the European Union of Geosciences, to be held in Nice, France, in April.

Most of these researchers eventually received their samples in December. For geochemist Erika Elswick and her colleagues at Indiana University in Bloomington, this left insufficient time to analyse sulphur isotopes in the boundary core and compare the results with cores from four sites across North America. From this work, Elswick had hoped to discover how far debris from the impact was thrown into the air. And besides being late, the boundary core samples were too small for the tests she needed to run. "We were dismayed," says Elswick. "There was no explanation given; no apology." Ames similarly received insufficient material for her planned geochemical studies, which aimed to identify ancient hydrothermal vents in the crater. Smit acknowledges that some researchers did not receive sufficient samples, but says that they didn't complain to him.

Gerta Keller, a palaeontologist at Princeton University in New Jersey, pressed Smit

hard to obtain samples for her studies of fossilized marine plankton known as foraminifera, most of which disappeared in the K/T extinction. Her analysis stunned the audience in Nice, suggesting that foraminifera continued to prosper for 300,000 years after the impact. Directly above the impact breccia — a layer of conglomerates that bears the marks of a powerful collision — Keller found foraminifera-rich sediments, which she dated using geochemical data and by analysing the microfossils they contained. The green line of the K/T boundary sat above these sediments. The Chicxulub impact, Keller concluded, could not have caused the mass extinction.

Core issues

Keller, who suspects that the K/T extinction was caused by a series of impacts, has long sparred with Smit, who is an enthusiastic supporter of the Chicxulub impact theory. She rejects his explanation for the delay in sending out the samples, claiming that "he tried to postpone our results so that he could remain unchallenged through the Nice meeting".

Smit dismisses that allegation as "ridiculous", and disputes Keller's scientific conclusions. The sediments in question, he argues, were not laid down over 300,000 years, but were instead washed back into the crater shortly after the impact. He also claims that Keller has misidentified non-living crystals as microfossils. But other researchers back Keller's identifications — Rebolledo has also conducted palaeomagnetic analyses of other sections of the boundary core and confirmed her dates.

Despite Smit's rejection of Keller's data, other experts believe that she has raised serious doubts about the link between the impact and the K/T extinction. "We ought to look very carefully at this," says Sharpton, who has previously supported the Chicxulub impact theory. "If she is right, we have a real issue."

The various groups on the Chicxulub project are now preparing articles for a special issue of the journal *Meteoritics & Planetary Science*, expected to appear next year. And from 30 September, researchers outside the project can submit requests to UNAM to analyse samples. Eventually, scientific consensus on whether or not the Chicxulub impact caused the K/T extinction may emerge. But the huge personal investment that many researchers in the field have put into their favoured theories is likely to ensure further acrimonious debate in the meantime. "It's not about science," claims Norman MacLeod, keeper of palaeontology at the Natural History Museum in London, who believes that sea-level changes and volcanism are more likely explanations for the K/T extinction. "It's about people's reputations."

Rex Dalton is *Nature's* US West coast correspondent.

♦ www.icdp-online.de/sites/chicxulub/news/news.html

Breeding to tackle blight without copper or GM

Gene mapping can help conventional breeders to focus efforts on building resistance.

Sir — Your well-aimed Editorial “Diversity in food technology” (*Nature* 424, 473; 2003), attacking self-damaging dogmas within the organic movement regarding the coexistence of ‘organic’ and conventional farming, cites the example of late potato blight caused by *Phytophthora infestans* and its control in organic farming by copper-based fungicides. Most potato cultivars (ancient and modern) could not be grown without some fungicide protection in either organic or conventional farming systems, although it is to be hoped that this will change. As things stand, between seven and twenty fungicide applications are made in a season to prevent crops from being destroyed by this devastating disease.

The use of copper compounds was scheduled to be prohibited across the European Union (EU) from March last year, but limited use has been permitted until 2006 in response to the needs of producers, mainly organic. The question is: why, under organic farming rules, are copper fungicides still allowed on potatoes, while much more modern, well-researched and safer fungicides are prohibited?

Copper-based fungicides are less effective against late blight and more toxic to non-target organisms than any of the more modern classes of anti-late-blight fungicide. They are also persistent in and

damaging to the environment, for example in soil. Extensive lobbying has gained organic producers the EU derogation for copper compounds, at a time when many other agrochemicals are being removed from the list of pesticides approved for use under the United Kingdom’s Control of Pesticides Regulations.

One can only surmise that copper is still supported by the organic movement because a copper-containing fungicide (‘Bordeaux’ mixture, invented by Pierre Millardet in France in 1882 for control of downy mildew on vines) is by far the oldest fungicide in regular use. It predates the organic movement by many years and is presumably venerated because it is ‘traditional’. But is tradition a good enough reason to continue using copper? Who is being truly organic and who conventional in defending this discredited nineteenth-century farming practice?

Although genetic modification (GM) technology has not provided a permanent answer to the worldwide scourge of late blight, it seems feasible that it may, given the intense research being undertaken around the world into the basic biology of the pathogen, the host and their interaction. Moreover, it is possible that such resistance will be based not on transgenes but on manipulation of existing potato genes.

That said, there is still much to be done

before that goal is reached and, contrary to the suggestion in your Editorial, much mileage in conventional breeding for high levels of late-blight resistance. The failure of earlier cultivars bred for high resistance to late blight was mainly due to the resistance being based on genes (R-genes) that provided resistance only to some strains of the pathogen. Resistance soon broke down when new strains arose through mutation or migration from elsewhere.

From the early 1960s, breeders have concentrated on sources of resistance that are not race-specific. Some of the genes involved have been mapped, and the nature and level of resistance that they contribute suggest that the resistance they confer will not break down rapidly. Some commercial varieties have high levels of late-blight resistance that has not yet broken down after years in cultivation, although there is no guarantee of immutability in biology.

It is perhaps ironic that one very recent variety, Lady Balfour, with a high level of late-blight resistance, has received an enthusiastic welcome, particularly from organic growers. Lady Balfour was of course one of the founders of the organic movement in the early twentieth century.

James M. Duncan

*Host-Parasite Co-evolution Programme,
Scottish Crop Research Institute,
Invergowrie, Dundee DD2 5DA, UK*

Despite Franklin’s work, Wilkins earned his Nobel

Sir — In this anniversary year of the publication of the classic paper proposing the double-helical structure for DNA (see www.nature.com/nature/dna50), renewed attention has been drawn to the respective roles of Rosalind Franklin and Maurice Wilkins in this ground-breaking discovery: see for example Watson Fuller’s Commentary “Who said ‘helix’?” (*Nature* 424, 876–878; 2003).

Some years ago, while working in the archives of the Pasteur Institute in Paris, I came across a document that may be of interest in relation to the appropriateness of the inclusion of Wilkins as a recipient of the 1962 Nobel prize for this work.

The document is a letter, dated 31 December 1961, and an accompanying overview of the DNA work, from Crick to his friend Jacques Monod, evidently intended to provide Monod with material for preparing a nomination letter for the Nobel. (Monod, who was to earn a Nobel

prize himself in 1965 for his work on gene regulation, was a senior researcher at the Pasteur Institute.)

Crick writes about Wilkins as follows: “On the matter of Maurice Wilkins. I think his contribution was two-fold. He initiated the careful X-ray work on DNA, and since 1953 he has done numerous extensive, accurate and painstaking studies on it. It is true that he has worked rather slowly, but then hardly anybody else has done anything. However, the data which really helped us to obtain the structure was mainly obtained by Rosalind Franklin, who died a few years ago. It should also be remembered that for a whole year Jim and I tried to get Maurice to solve the structure by our approach, without success. It was only after we learnt of Pauling’s structure that we asked and obtained Maurice’s permission to work on the problem. Nevertheless, for the last eight years Maurice has done all the hard work on the problem and that should be recognized.”

Although he thus clearly gives priority to Franklin, Crick in his overview credits

Wilkins with initiating “the only serious X-ray work on DNA up to 1953”, with being “the first person to realize that DNA might be helical”, and with carrying out the bulk of the work on DNA from the time that Franklin left King’s College to work with John Bernal at Birkbeck.

Judging from the handwritten revisions and editorial changes made by Monod, which can be clearly seen, it seems that Monod drew directly from Crick’s information when he wrote his letter of nomination. It also seems that Crick himself thought that the significant body of work produced by Wilkins — before and after Franklin’s crucial contribution — merited his inclusion in the nomination.

Doris T. Zallen

*Department of Science and Technology
in Society, Virginia Tech, Blacksburg,
Virginia 24061-0247, USA*

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Learning to teach

Can studies of how people learn be put to use in the classroom?

Original Intelligence: Unlocking the Mystery of Who We Are

by David & Ann Premack

McGraw-Hill: 2003. 274 pp. \$24.95

Peter Bryant

Cognitive scientists study the nature of intelligence, and therefore of learning, and several of them have in recent years turned their attention to children's education. The reason is simple and impeccable: they have made discoveries about how people acquire knowledge that should be of value to schoolteachers in helping children learn new things. This is the approach taken by David and Ann Premack in this highly impressive book on intelligence and education.

The Premacks start their book with an account of the basic mechanisms of learning, and end it with a set of specific suggestions for ways of teaching children in schools. In between they set out their conclusions about their own research and that of others on topics of central importance in cognitive psychology, including language, mathematics, memory, causal reasoning and reasoning through analogy.

In their opening remarks on learning, the Premacks argue strongly for the existence and importance of modules — highly specialized mechanisms for acquiring different forms of knowledge. According to modular theories, children largely depend on modules for learning to speak (the language module), for finding out how to count, add and subtract (the number module), and for acquiring everyday biological and physical knowledge (the biological and physical modules). The Premacks also review some of the physiological support for the idea of separate mechanisms for learning. Modular theories, it should be said, resonate well with the geographical approach that has dominated cognitive neuroscience for several decades now — the idea that this bit of the brain underlies our spatial knowledge, another bit over there subserves language, and right at the front somewhere you will find the regions responsible for planning, say, and perhaps theory of mind.

The chapters that follow deal with topics such as analogy, and causal and social reasoning, which are at the centre of the Premacks' justly famous research using chimpanzees and, later on, young children. The ingenuity and originality of these experiments are truly astonishing, and the authors' account of the work is clear and exciting. It is also impressive to see how many new ideas the Premacks come up with in these chapters. They challenge, for instance, the traditional idea that



Waiting for the penny to drop: cognitive science may be able to help children learn.

belief is an intervening variable. It is often held that people build up beliefs through experience and inferences, and then act on these beliefs. But on the basis mainly of work with young children, the Premacks argue instead for a direct route between our perceptions and inferences and what we do.

The ideas in the chapter on the theory of mind are also strikingly original. This topic is about our ability to work out what is in another person's mind. The Premacks introduced this hugely important issue into cognitive psychology in the 1970s, and it would have been easy for them to spend this part of the book resting on their well-deserved laurels. Instead, they quite rightly criticize the way in which success in 'false belief' tasks has become the modern gold standard for holding a theory of mind. These tasks test whether a child can understand that someone else can hold an incorrect but quite legitimate belief about an object or a set of events. Even though some of their own initial research on theory of mind involved false beliefs, the Premacks argue here that other criteria and tasks are just as important, and give a stimulating account of some experiments on understanding the intentions of others engaged in goal-directed actions.

Early on in the book there is a chapter on pedagogy, in which the Premacks claim that "the root of human pedagogy lies in aesthetics, in the judgement of beauty which, like charity, begins at home". I find this particular argument hard to follow, and hard to link to

their correct assertion in the same chapter that humans need teachers more than other species do because specialization in human intelligence has led to inventions and intellectual advances that we cannot expect successive generations to learn without help. These human achievements, which L. S. Vygotsky called 'cultural tools', have to be taught, and at the end of their book the Premacks turn their attention to how best to do this teaching.

The journey from the psychological laboratory to the school classroom is complicated and hazardous, for these are two very different places. "You make a great, a very great mistake," the psychologist William James once thundered, "if you think that psychology, being the science of the mind's laws, is something from which you can deduce definite programmes and schemes and methods of instruction for immediate schoolroom use." If this is a mistake, then I am afraid that the Premacks seem to have made it. They outline educational programmes but provide no evidence that they actually work. Such research is difficult to do — difficult, but essential.

The Premacks also ignore a large amount of existing research, both psychological and educational, on how to help children make the advances that they themselves wish to encourage. For instance, their claim that "the basic idea of a fraction can be taught by dividing an object into a number of parts" actually describes a fairly widespread classroom practice, and sidesteps some excellent research on children's learning about proportions and fractions that certainly deserves a mention. The Premacks' assertion that children learning to read will come to segment words into various phonological units by seeing their own speech transformed into print is interesting and worth a try. However, a great deal of quite sophisticated work on children's awareness of phonology suggests that it won't work.

Why wasn't any of this work mentioned in this book? I think the reason is that the Premacks' main educational idea is that one ought to give modules, which are innate learning mechanisms, a chance to learn. The twists and turns of research into children's developing understanding of proportions, phonology or physics do not fit the Premacks' bill, which is a pity because not all of this research is bad. This educational chapter provides an interesting but downbeat ending to a provocative and valuable book. ■

Peter Bryant is in the Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK.

Troubled waters

FitzRoy: The Remarkable Story of Darwin's Captain and the Invention of the Weather Forecast
by John and Mary Gribbin
Review: 2003. 336 pp. £18.99

Evolution's Captain: The Tragic Fate of Robert FitzRoy, the Man who Sailed Charles Darwin Around the World

by Peter Nichols

Profile Books: 2003. 352 pp. £16.99
HarperCollins: 2003. 352 pp. \$24.95

Keith Thomson

Robert FitzRoy was one of the most gifted naval officers of his or any generation, a brilliant seaman, surveyor and commander of men, and a technical innovator who in different circumstances might have been a great scientist. Most of the characters who entered our lives through the voyage of the *Beagle*, which FitzRoy captained, went on to become part of a great Victorian success story. FitzRoy did not.

Aristocratic, handsome and wealthy, at times he could be carefree and capable of achieving anything he set his mind to. The few surviving letters between him and Charles Darwin from the voyage of the *Beagle* reveal an almost giddy schoolboy companionship that is hard to connect to his episodes of outrageous behaviour and anger (over slavery, for example). He was also a workaholic, dogged both by his own depressive mental illness and by the knowledge that others in his family had suffered the same way.

His life was full of tragedies, usually generated by his own mercurial temper, sudden black moods and constant brooding unhappiness, compounded by official resistance to his cutting corners and, in his paranoia, the effortless successes of others. After the exciting years on the *Beagle*, he never attained true greatness, whether as an officer (he never had another proper command), as a member of parliament (for two years), as a brave but unappreciated governor of New Zealand (he was recalled for siding with the Maoris), or as a meteorologist. When all of the pressures overcame him, he copied his uncle Lord Castlereagh, and cut his own throat with a razor.

Somewhere, it is to be hoped, there exist the letters and diaries that will show us a fuller version of the life of this strangely brilliant man, beyond the stereotypes — a different FitzRoy from the one who seems always to be in Darwin's shadow. At a critical time, FitzRoy must have known Darwin better than anyone else, but after a hundred or more years of the 'Darwin industry', we still do not know what, if anything, was FitzRoy's contribution to Darwin's



Looking ahead: Robert FitzRoy, the *Beagle's* captain, went on to invent weather forecasting.

intellectual development. For example, it is hard to guess how different Darwin's later career would have been if James Sullivan (mate on the *Beagle*, who was later knighted and made an admiral) had been captain. Nor do we know how the voyage with Darwin influenced FitzRoy's later career. We know only that they shared meals, not a philosophy.

The Gribbins guess that FitzRoy became a religious fundamentalist as a result of spending long, lonely hours studying the Bible during the later parts of the voyage. But all we know for sure about this is that Darwin was not an evolutionist until after the voyage, and that FitzRoy was not a fundamentalist until his marriage to the religious Mary Henrietta O'Brien in 1836.

To the only previous biography of FitzRoy (*FitzRoy of the Beagle* by H. E. L. Mellersh; Hart-Davis, 1968), John and Mary Gribbin have added information from the few new letters to come to light. Peter Nichols' book, which has too many irritating errors, diversions into side issues and purple passages, and an absence of either citations or index, is less useful. Both books devote too much space to rehashing the *Beagle* voyages — something that has been covered so much better elsewhere. It is in FitzRoy's life after the *Beagle* that his contributions to science and modern life are to be found. FitzRoy the unusually enlightened Victorian public servant, or the man who gave us synoptic weather charts and almost single-handedly invented weather forecasting, is at least as interesting as the one who, as Nichols dismissively puts it, "sailed Darwin around the world".

The Gribbins develop the idea that FitzRoy, who was probably always a manic-depressive and who unthinkingly gave more of himself and his fortune than was wise,

committed suicide in part from guilt at having helped the birth of the darwinism 'heresy'. If that's the case, we would need to have evidence of what FitzRoy thought his contribution had been. He was also depressed about the US Civil War and the situation of Matthew Maury, an exile in Paris who also had a claim to have invented weather forecasting. Almost bankrupt himself, FitzRoy still sent his rival money to live on. And the great 'machine' that supported and promoted Darwin — in this case principally in the form of Francis Galton — never hesitated to stick the knife in.

But there must have been much more to FitzRoy than that. As the Gribbins lament, however, when it comes to FitzRoy we know everything in terms of names, dates and places but very little about the man. FitzRoy was one of the most compelling figures of the darwinian era, but his fate has been to remain hidden. We want to know more and we fervently hope that it might include a little more triumph to go with the tragedy. ■

Keith Thomson is in the Oxford University Museum of Natural History, 1 Parks Road, Oxford OX1 3PW, UK.

The universal engineer

Discussion of the Method: Conducting the Engineer's Approach to Problem Solving
by Billy Vaughn Koen

Oxford University Press: 2003. 272 pp.
\$60, £45

Julio M. Ottino

This book is about the universal method. "Philosophers have long sought the universal method... They failed to look for it in the most unlikely of places, in the engineering method," says Billy Vaughn Koen, in his book *Discussion of the Method*. An engineer solves problems using heuristics — anything (such as state-of-the-art science, present and past technology, or just plain experience) that provides a plausible aid or direction in the solution of a problem but is, in the final analysis, unjustified, incapable of being justified, and fallible. When faced with a situation that is complex and poorly understood, and with limited resources at your disposal, if you seek a solution that is the best available, then you too are an engineer. "To be human is to be an engineer," as Koen puts it.

In Koen's view, engineering is often confused with its products, be they antiquated irrigation channels, aqueducts, gears and bearings, or the more recent computers, chemical plants and aeroplanes. In this book he argues for a wider unifying view based

on methodology — an “engineering way of thinking”. This provocative book is full of ideas with heuristics as the centrepiece. Heuristics appear in many disciplines, from computer science and artificial intelligence to operations research, but rarely are they elevated to the exalted position they occupy in this book.

In the second part of the book, the author puts the case for expanding the engineering method into a ‘universal method’. The book touches on so many topics that it is hard to imagine philosophers, mathematicians, linguists, physicists and scientists of all stripes not having an opinion about it.

At the centre are heuristics as an incremental and sure-footed driver of change. According to Koen, “the engineering method is the strategy for causing the best change in a poorly understood situation within the available resources”. But clearly not all changes are equal, and some are certainly not sure-footed; some jumps seemingly come out of nowhere. Is this a claim of universality going too far? Is heuristics — from the Greek *heuriskein*, meaning to discover — a path to creation and invention as well? I’m not convinced, but Koen is unequivocal: all is heuristic.

Technology is about invention, making and building, which can be rather haphazard processes. The appearance of the video format VHS was not inevitable — we could have jumped straight to DVDs, or passed from Betamax to DVDs. The process is like playing with an ever-growing Lego set — some pieces are already there, and combinations of these lead to the creation of new pieces, which lead to yet more. There is no going back. But one can see how a growing heuristic may actually guide the process.

But can the processes of creation and discovery be viewed as a part of a heuristic? Immanuel Kant, stating his position on creation and its uniqueness, declared that science is ephemeral, and that art is permanent. The painting *Les Femmes d'Alger* would not exist without Picasso. Not everybody agrees. The art of writing, Jorge Luis Borges argued, is actually discovery, revealing what was already there. Likewise, science can be viewed as discovery, and hence as an inevitable process: Newton’s laws would have been revealed eventually, although possibly in a different form to that obtained by Newton. Parallel universes with the same physics and chemistry will have the same science, but may have different technologies.

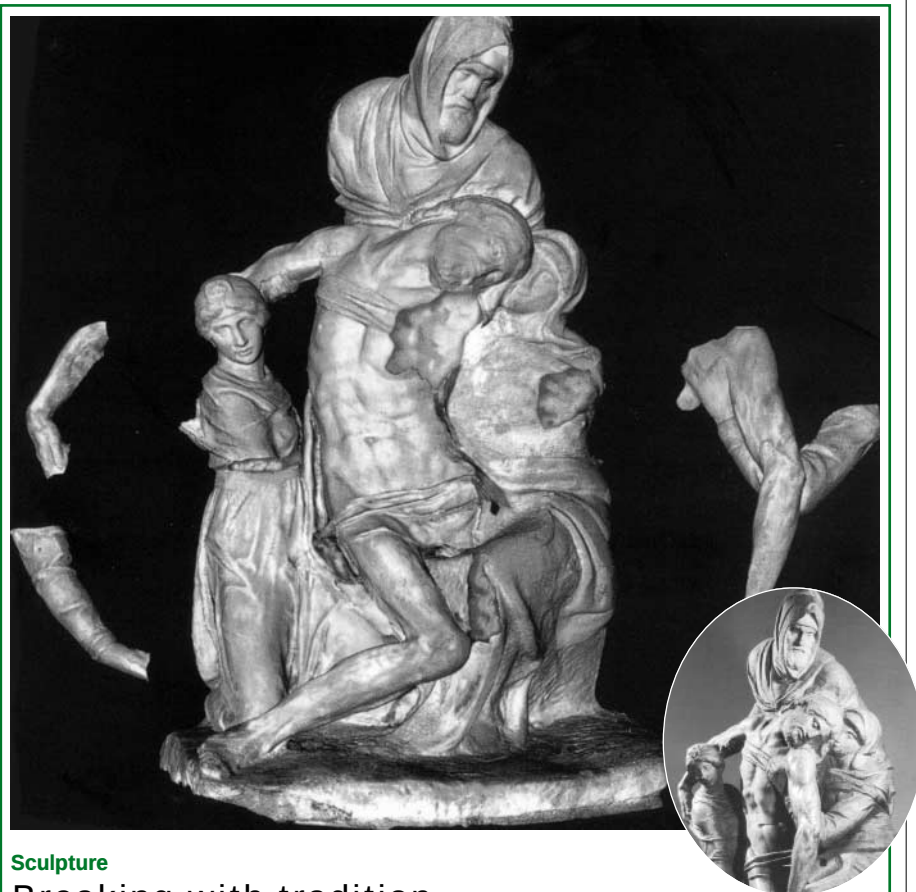
But there is no inevitability in art. The history of modern art appears to be driven by replacing and disavowing heuristics. If any heuristic exists, it is one of replacing heuristics. One cannot help but wonder whether Koen’s claim is simply too big. There are ways out: we are told that a heuristic may contradict other heuristics, and that use may depend on context. And even

though the author takes pains to indicate that science is part of the heuristics, some who read the book superficially will equate engineering with a much narrower definition of heuristics.

Even if one only partly agrees with the main conclusions of the book, many of the arguments are thought-provoking. Break-throughs can be seen as ‘break-withs’ the prevailing state-of-the-art. And the idea that a heuristic-based approach is much older than the scientific method is intriguing to say the least.

Is the method proposed in this book a good way to think about the crucial issues facing us today? One can think of many cases where one wishes that this mode of thinking had been followed. The first part of the book should be required reading in science and engineering departments. It has the potential to create a huge controversy — what it doesn’t deserve is indifference. ■

Julio M. Ottino is in the R. R. McCormick School of Engineering and Applied Sciences, Northwestern University, Evanston, Illinois 60208, USA.



Sculpture

Breaking with tradition

Michelangelo's sixteenth-century *Pietà* in the Museo dell'Opera del Duomo in Florence, Italy, incorporates a mystery as deep as its expressive beauty. Having worked for a decade on the carving he intended for his own tomb, Michelangelo apparently attacked his 2.5-metre sculpture in a fit of rage, hacking limbs from the four carved figures. These were later re-attached by his assistant, apart from Christ's left leg, which remains missing.

Why did he do it? Did he fear it fell short of perfection? Was he angered by a flaw emerging in the marble? Had he been afraid that his positioning of the left leg across the Virgin's lap was immodest?

In a collaborative project with art historian Jack Wasserman, scientists have used IBM's most sophisticated three-dimensional imaging

technologies to create a 'virtual *Pietà*'. This has enabled Wasserman to learn more about how the sculpture was physically carved and repaired. It has also allowed him to put forward his theory that Michelangelo was not intending to destroy his work in anger at all. By 'virtually dismembering' the re-attached limbs, and studying the carving from all angles, Wasserman concludes that the limbs had been detached selectively, and with care, perhaps to create a composition where Christ is being lowered from the Virgin's lap, rather than into it.

Wasserman's book *Michelangelo's Florence Pietà*, published by Princeton University Press (\$75), describes the extraordinary technical analysis of one of the most important artworks of post-Renaissance Italy. It also includes a CD of the virtual *Pietà*.

Alison Abbott

Flexible arrangement

Neil D. Theise and Ian Wilmut

Until recently it was generally thought that cells move forwards along their respective differentiation paths, but never backwards, and certainly do not jump from one path to another. This dogma of unidirectional, hierarchical cell lineages in tissue development, maintenance and repair is explained by the action of irreversible gene restrictions. As cells differentiate in a lineage, genes that might be required for other pathways are irreversibly repressed.

However, exceptions to the loss of plasticity associated with such lineage restrictions have long been recognized in disease and repair. For example, the epithelium that lines the lungs of smokers is often seen to change from simple columnar cells to a stratified configuration (a process called squamous metaplasia), and bone can be formed in injured skeletal muscle (osseous metaplasia). Experimentally, heterokaryons, which are created by transferring a nucleus from a cell of one type into a cell of a different type, show changes in nuclear gene expression that reflect the character of the host cell, demonstrating that differentiation is an actively maintained, dynamic state rather than a one-way street.

With the blossoming of stem-cell research, demonstrations of heretofore implausible genomic plasticity are now published almost weekly. Many reports describe the derivation of cells of several tissues from a single source population. Although the mechanisms of genomic plasticity remain poorly understood, the presence of plasticity suggests that gene-restriction mechanisms are not irreversible after all.

Four plasticity pathways have been documented *in vivo* and experimentally. These pathways may involve undifferentiated cells, situated within specialized tissues, that can switch developmental programmes in response to injury. In the liver, for example, the tiniest cells lining the bile duct are

'bipotent' — they can regenerate either hepatocytes or other, larger bile-duct-lining cells in the face of injury. If tissues contain truly totipotent cells, these cells might be 'embryonic rests' persisting in adult tissues long after embryonic development is completed. Another possibility is that differentiated cell types may 'de-differentiate' to an earlier, progenitor phenotype. This process is probably more common in neoplasia, particularly malignancy. Alternatively, differentiative leaps can be induced by experimental manipulation or, *in vivo*, in response to injury. Thus, cells of differentiated phenotypes can have wide developmental ranges, and are not confined to the tissues from which they are derived. In such 'transdifferentiation' events, the influence of microenvironment, perhaps through changes in response to injury, would be key. Finally, fusion between cells in some injury models can lead to reprogramming of nuclei, similar to that seen in *in vitro* heterokaryon experiments.

The most dramatic demonstrations of nuclear plasticity have been provided by the birth of offspring after the transfer of nuclei from adult somatic cells — the most famous case being Dolly the sheep. This process depends on the reprogramming of the transplanted nucleus by factors in the egg's cytoplasm. Although offspring have so far been obtained with a variety of donor cells in seven mammalian species, the process is very inefficient: less than 5% of embryos survive to adulthood, signifying a failure to 'reset' most somatic nuclei. Oocytes evolved to bring together nuclei packed in oocyte and sperm proteins, respectively, and to establish appropriate chromatin structure for normal development. It is no surprise, then, that their cytoplasm is unable to remodel efficiently the transferred nuclei that are organized for an alternative pattern of gene expression.

However, it is clear that at least two nuclear transcription factors, called Oct-4 and Nanog, have some ability to restore embryonic-like plasticity to mature adult cells. It is likely that at least some of the cytoplasmic reconditioning of nuclei — as seen in experiments involving cloning, heterokaryon/cell fusion, or reprogramming by administration of cytoplasmic extracts — may work through induction of these nuclear factors.

Molecular mechanisms of gene repression are the barriers over which adult cells must leap to become plastic. Such repression can arise from direct molecular modifications of DNA, such as methylation of cytosine residues in clustered C–G pairs near promoter sites, which usually suppresses the associated gene. It can also result from methylation and/or deacetylation of histone proteins,

Cell plasticity

Elucidation of all of the mechanisms that regulate developmental potential will allow us to discover the true limits of cell differentiation.

around which the DNA is coiled, forming 'heterochromatin' regions that are unavailable for transcription. Furthermore, such regions are often topographically located at points of attachment to other structures within the nucleus, providing exceptionally stable, if not actually rigid, three-dimensional structures. This conformation then leaves other regions (euchromatin) flexible and exposed to factors that can initiate transcription.

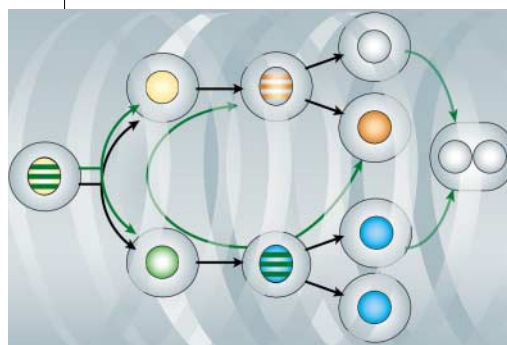
While these mechanisms of gene restriction serve to explain cell lineages that appear, at the grossest level of examination, to be unidirectional and hierarchical, several researchers are now demonstrating physiological mechanisms for their reversal. For example, passive reversal of DNA methylation can occur during gene replication when cytosine residues of the newly formed DNA strand are left unmethylated. Methylation must also therefore be actively maintained; if it is not, de-repression will occur. Demethylases are also responsible for the active demethylation of both methylated cytosines and some methylated histones. Tissue- and cell-specific histone acetyltransferases have been identified — these molecules allow heterochromatin and euchromatin domains to shift, resulting in de-repression of tissue-specific genes.

Elucidation of all of the mechanisms that regulate developmental potential will allow us to discover the true limits of cell plasticity. Progress will depend upon studies and manipulation of both the extra- and intracellular factors that influence the fate of cells. Furthermore, exploration of likely overlapping mechanisms between genetic de-repression in cloning and adult-plasticity experiments, as well as the mechanisms that maintain unexpressed gene states in embryonic stem cells, may be particularly fruitful. The rewards for this research will be a far better understanding of developmental mechanisms, and important new opportunities to derive cells of specific phenotypes for research and medicine. ■

Neil D. Theise is in the Division of Digestive Diseases, Beth Israel Medical Center, New York 10003, USA. Ian Wilmut is at the Roslin Institute, Roslin BioCentre, Midlothian EH25 9PS, UK.

FURTHER READING

Blau, H. M. *Nature* **419**, 437 (2002).
Wilmut, I. *et al.* *Nature* **419**, 583–586 (2002).
Theise, N. D. & Krause, D. S. *Leukemia* **16**, 542–548 (2002).
Cavaleri, F. & Scholer, H. R. *Cell* **113**, 551–552 (2003).



The winding road: cell differentiation involves more than a simple one-way progression.

Tracking the first Americans

Tom D. Dillehay

A study of 33 ancient skulls excavated from Mexico invites us to reconsider our view of the ancestry of the early Americans. Unlike most other early American remains, the skulls resemble those from south Asian populations.

Questions of which human populations first arrived in the Americas, and when, where and how this happened, have been debated by scientists for decades¹. It has long been presumed that the first people entering the New World were the direct ancestors of present-day Native Americans and that they arrived in America from northeast Asia about 12,000 years ago². But this theory has been challenged by new archaeological discoveries and by findings of early human remains bearing anatomical similarities to the people of south Asia and the southern Pacific Rim^{3,4} (Fig. 1). Writing on page 62 of this issue, González-José *et al.*⁵ add more fuel to this heated debate. They present a comparative study of early historic human skulls from Baja California, Mexico, and their findings lend weight to the view that not all early American populations were directly related to present-day Native Americans.

Human skeletal remains have long been used by palaeoanthropologists to model early human migration. The conventional view is that different skeletal populations with similar craniofacial features (skull form) shared a common ancestry and were genetically related, whereas different features reflect

different ancestry. Migration histories and evolutionary forces explain the similarities or differences.

Piecing together the ancestry of the Americas has been difficult, as early human remains dating from about 10,000 years ago (the end of the last ice age) are fragmentary and scarce. Scientists have typically reconstructed the missing pieces of the most ancient skulls by extrapolating backwards from later, more complete skeletons. Ancient American skulls reconstructed in this way were anatomically indistinguishable from early northeast Asians and also from present-day Native Americans². So a theory arose, supported by dental and other archaeological data⁶, that the first humans entering the Americas were northeast Asians who arrived in three successive migrations beginning around 12,000 years ago. These founding colonizers were thought to be big-game hunters, equipped with so-called Clovis spears⁷, who rapidly populated the Western Hemisphere and gave rise to present-day Native Americans (Fig. 2a).

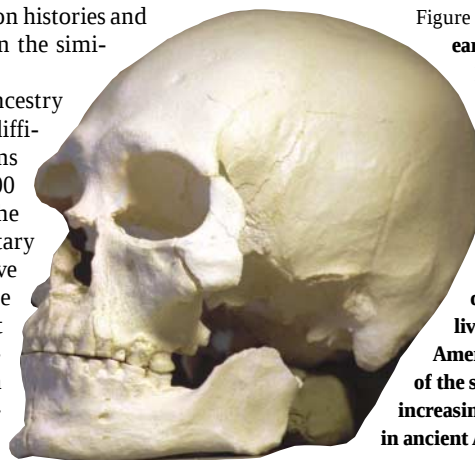


Figure 1 The skull of an early American. The Kennewick Man, excavated from Washington state, lived about 9,300 years ago. His craniofacial features, which do not resemble living Native Americans, are typical of the skulls that are increasingly being found in ancient American sites.

But more recent archaeological discoveries suggest that there were several different founding populations, arriving from different places, each with different lifestyles and technologies⁸. Some populations not only hunted big game but also exploited a wide range of plant and animal life. To complicate matters further, it is no longer certain that the first colonizers arrived about 12,000 years ago — some archaeological sites in South America date from 12,500 years ago, which suggests that the first humans arrived at least 15,000 years ago.

A similar pattern of diversity is emerging from statistical analyses of cranial and facial measurements of some of the oldest skeletons found in the Americas. The archaeological and skeletal data have led to a new model, in which the Palaeoamericans — the proposed first arrivals in the New World — were not northeast Asians. They came instead from south Asia and the southern Pacific Rim, and they probably shared ancestry with ancient Australians and other southern populations^{3,9}. A second group of humans then arrived from northeast Asia or Mongolia, and it was this second population that adapted to the warming climate after the Ice Age and gave rise to the modern Amerindians (an ancient population of Americans whose skeletal remains make up most of the human material found in the New World) and the present-day Native Americans (Fig. 2b). So according to this theory, the Palaeoamericans are unrelated to most modern Amerindians and to the Native Americans.

González-José *et al.*⁵ now propose a more complex view of American ancestry.

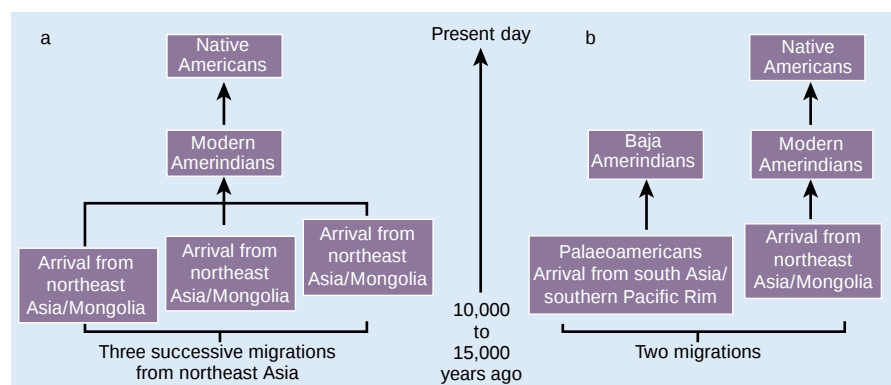


Figure 2 Tracing American ancestry. Analysis of skeletal remains has led to at least two models to explain the origin of early human populations in the Americas. a, It was originally thought that the first colonizers were the direct ancestors of present-day Native Americans, who arrived from northeast Asia and possibly central Asia in three successive migrations about 12,000 years ago. b, More recent analyses of the craniofacial features of skulls dating from the end of the Ice Age suggest that the first arrivals were from south Asia or the Pacific Rim. These 'Palaeoamericans' were thought to be unrelated to the majority of modern Amerindian remains — a later group of colonizers from northeast Asia were thought to have given rise to these late-prehistoric populations. Now, González-José *et al.*⁵ have found that a group of early historic Amerindian skulls from the Baja peninsula in Mexico bear a strong resemblance to the early Palaeoamericans, suggesting that the colonization of the Americas was more complex than had previously been suspected.

These authors analysed the skeletal remains of 33 modern Amerindians from early historic times, excavated from the tip of the Baja peninsula in Mexico. Surprisingly, the craniofacial features of these Baja Amerindians show closer affinity to the Palaeoamerican skulls than to other modern Amerindian remains. The Baja Amerindian and Palaeoamerican skulls have similar long and narrow braincases and relatively short, narrow faces, implying a common ancestry with the ancient inhabitants of south Asia and the Pacific Rim. González-José *et al.* confirm that modern Amerindian skulls from other areas are similar to ancient northeast Asian remains. Their new data add to accumulating evidence of morphological differences between early humans from different areas of the Americas^{8,9}.

The authors consider several potential explanations to account for the presence of Palaeoamerican traits in the Baja Amerindian skulls, but they suggest that the best explanation is that the Palaeoamericans were the direct ancestors of the Baja Amerindians. After the Ice Age, the increased aridity could have geographically isolated the founding Palaeoamerican population in the Baja area, and limited its gene flow with other modern Amerindian groups.

Do the new findings tell us anything more about when the first humans arrived in the Americas? The authors do not fully discuss the chronological implications of their work, but their interpretation of shared ancestry between the Palaeoamericans and the Baja Amerindians might best fit a model of Palaeoamerican arrival about 11,000–12,000 years ago. There is no direct evidence to support this view, but if the Palaeoamericans had arrived 15,000 years ago or earlier, the Baja population would have remained isolated for much longer. This seems unlikely, given the rate of population growth and movement that probably occurred after initial colonization and then after the Ice Age when the climate warmed.

But could the similarities between the ancient Palaeoamericans and the later Baja Amerindians instead reflect the influence of other evolutionary forces, such as gene flow or natural selection and convergent adaptation of different populations to similar local environments? Answering this question will depend upon finding more isolated prehistoric populations showing ancient Palaeoamerican traits, and then establishing whether parallel evolutionary forces were acting on them and whether they were derived from a single ancestry. But this will be a difficult task. Human remains from the end of the Ice Age are scarce and often fragmentary, so we have only a vague notion of the skeletal characteristics of the ancient Palaeoamericans. And we have a poor

understanding of the migration history of different American populations and what kind of evolutionary forces might have influenced them¹⁰.

Given these limitations, the findings of González-José *et al.* do not allow us to draw firm conclusions about the relationship between the ancient Palaeoamericans and the later Baja Amerindians. But the importance of this and other studies^{7,11} is that they suggest a different view of the origins and interactions of early human populations in the Americas. What we really want to know is what took place within and between these populations, how they changed over time, and how quickly they changed. These issues can be resolved only by obtaining more skeletal data¹¹ and by combining them with regional archaeological records, which should provide information on the social and cultural histories of the different

populations. Slowly, we are realizing that the ancestry of the Americas is as complex and as difficult to trace as that of other human lineages around the world. ■

Tom D. Dillehay is in the Department of Anthropology, University of Kentucky, Lexington, Kentucky 40506, USA.

e-mail: dilleha@uky.edu

1. Dixon, E. J. *Quat. Sci. Rev.* **20**, 277–299 (2001).
2. Dalton, R. *Nature* **422**, 10–12 (2002).
3. Neves, W. A. & Pucciarelli, H. M. J. *Hum. Evol.* **21**, 261–273 (1991).
4. Dillehay, T. D. *The Settlement of the Americas: A New Prehistory* (Basic Books, New York, 2001).
5. González-José, R. *et al.* *Nature* **425**, 62–65 (2003).
6. Turner, C. G. *Mem. Calif. Acad. Sci.* **27**, 123–158 (2002).
7. Meltzer, D. J. *Mem. Calif. Acad. Sci.* **27**, 27–58 (2002).
8. Lahr, M. M. *The Evolution of Modern Human Diversity* (Cambridge Univ. Press, 1998).
9. Neves, W. A. *et al.* *Homo* **50**, 258–263 (1999).
10. Steele, D. G. & Powell, J. F. *Mem. Calif. Acad. Sci.* **27**, 93–122 (2002).
11. Jantz, R. E. & Owsley, D. W. *Nevada Hist. Soc. Q.* **40**, 62–84 (1997).

Earth science

Just add water

Albrecht W. Hofmann

A new model could explain why Earth's upper mantle is depleted of many trace elements. At a certain depth, minerals might release water, creating a molten filter that traps trace elements in the mantle beneath.

Sandwiched between Earth's thin crust and its metallic core lies a layer of pressurized rock at high temperature — the mantle. Convection in this layer drives plate tectonics and sea-floor spreading, but we know little about the pattern of circulation. Indeed, current thinking about mantle dynamics is in a state of turmoil. As we cannot observe convection directly, we must piece together indirect evidence from seismology, geochemistry, mineral physics, fluid dynamics and numerical simulations of convection. But the evidence is contradictory and has led to at least two conflicting views about mantle movement. On page 39 of this issue, Bercovici and Karato¹ propose a new model that might resolve this conflict. They suggest that water dissolved in the mantle might create a thin layer of melt at a depth of 400 km, causing an unexpected pattern of circulation in the mantle.

The two conflicting models for mantle convection (Fig. 1a, b, overleaf) are usually described as 'layered' convection (supported by geochemists) and 'whole-mantle' convection (supported by seismologists). Geochemists have long insisted on the two-layered model, in which the mantle consists of a relatively primitive layer below a depth of 660 km — containing primordial noble gases, trapped 4.5 billion years ago when the Earth formed — and an upper layer that is highly depleted of heat-producing elements (uranium, thorium and potassium), noble

gases and other 'incompatible' elements. The primitive layer serves as a reservoir for these elements (which were depleted from the upper mantle when Earth's crust was formed) and it is occasionally sampled by deep-mantle plumes (Fig. 1a).

Over the past several years, however, seismic tomography has given us increasingly detailed images of apparently cold 'slabs' (characterized by fast seismic velocities) descending into the deep mantle right through the 660-km boundary, effectively cutting to shreds the simple picture of the mantle convecting in two nearly isolated layers. If cold 'slabs' descend into the deep mantle, there must be a corresponding upward flow of deep-mantle material to shallow levels (Fig. 1b). No matter what specific form the exchange across the 660-km boundary takes, in this 'whole-mantle' model of mantle convection, it would within a few hundred million years destroy any compositional layering that had possibly been inherited from early in Earth's history.

Meanwhile, the geochemical arguments for a separate deep reservoir have not disappeared. Primordial noble gases are still preferentially associated with 'hot spots'², at least some of which seem to come from deep-mantle plumes³. And much of the upper mantle remains highly depleted of incompatible trace elements, including the heat-producing thorium, uranium and potassium — also suggesting the presence of a less

depleted reservoir deep within the mantle⁴.

Can these conflicting data be reconciled? Some scientists have proposed that a relatively small, seismically 'invisible' primitive reservoir, which is slightly denser than ordinary mantle material, is trapped in the deep mantle^{5,6}. Others suggest that the seismically anomalous, so-called D'' layer just above the core–mantle boundary can satisfy the geochemical requirements⁷. Yet another theory is that the geochemical heterogeneities are dispersed throughout the mantle but are preferentially sampled by hot-spot magmas, which therefore deliver a biased record of their source chemistry⁸.

Bercovici and Karato¹ propose a new model for mantle convection that might reconcile the geochemists and the seismologists. Experimental work has shown that, at high pressure, the mantle minerals wadsleyite and ringwoodite have surprisingly high solubilities for water⁹. High-pressure forms of these minerals occur in the depth range of 410–660 km — the so-called mantle transition zone — and Bercovici and Karato postulate that this zone has a high water content (0.2–2.0% by weight). They suggest that when material in the transition zone rises through slow convective motion to depths of less than 410 km, wadsleyite is transformed into the mineral olivine and liberates water (olivine has a much lower solubility for water than wadsleyite).

Adding water to almost any rock will lower its melting point by several hundred degrees. The authors suggest that this 'dehydration melting' might create a thin (about 10-km) layer of molten silicate just above the 410-km level, which would retain most of the water as well as all those incompatible trace elements (including uranium and thorium) that do not easily fit into the crystal structures of the available solid minerals. The melt layer is trapped at 410 km, because melt is denser than solid at very high pressures, but it is ultimately entrained, refrozen and returned to the transition zone by descending cold, subducted slabs. This return flow keeps the thickness of the molten silicate approximately constant. So the melt layer acts as a filter, removing elements from rising mantle material and keeping the upper layer in a depleted state (Fig. 1c). Deep-mantle plumes are exempt from this filtering process because their higher temperature lowers the solubility for water in wadsleyite, but increases solubility in olivine, thus preventing the formation of a melt layer at a depth of 410 km.

The new model has not yet been tested by numerical simulations of mantle convection. But this might prove to be difficult or impossible to do in the near future because the melt layer is very thin and its mechanical behaviour is complex. But does the model satisfy the geochemical constraints? Many isotope geochemists will say that it does not,

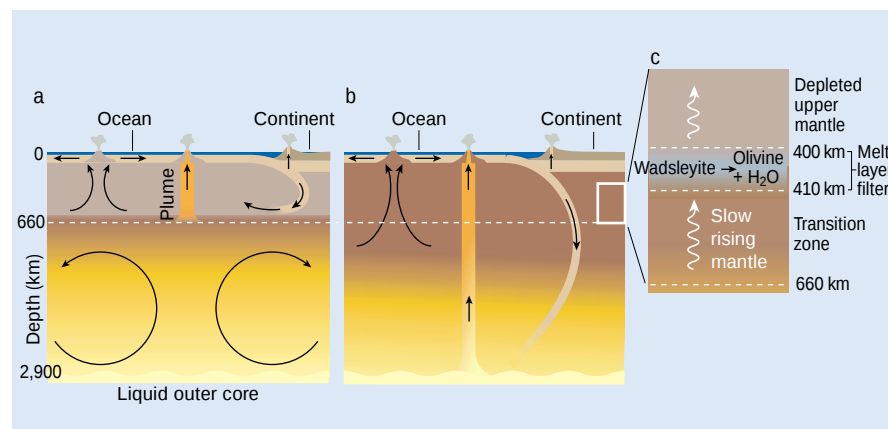


Figure 1 Models of mantle circulation. a, Geochemists support a two-layered model in which the upper layer of mantle, depleted in 'incompatible' elements, is separated from the primitive lower mantle by a boundary at a depth of 660 km. b, Seismologists disagree. They support a 'whole-mantle' model of circulation with exchange between subducting crust and upwelling plumes. c, Bercovici and Karato¹ now propose a model that might satisfy both camps. They suggest that at depths of 410 km the mineral wadsleyite releases water and is transformed into olivine. This creates a thin layer of molten silicate that acts as a filter, removing 'incompatible' trace elements from slowly uprising mantle.

but I am not so sure. The difficulty comes in reconciling Bercovici and Karato's hypothesis with our view of how the mid-ocean ridge forms. This part of Earth's crust is thought to be derived from the upper mantle, whereas ocean islands are made from deeper mantle plumes. These different source materials can be traced by the composition of strontium isotopes in the basalt rock — mid-ocean ridge basalts have lower ⁸⁷Sr/⁸⁶Sr ratios than ocean island basalts. The ⁸⁷Sr isotope is created by the very slow decay of ⁸⁷Rb (with a half-life of about 47 billion years), so the isotopic differences between the upper mantle and the deeper mantle take a very long time to develop, typically 1–2 billion years.

In the Bercovici–Karato model, the upper mantle is continuously replenished by slowly rising deep-mantle material. The melt layer at 410 km removes water and incompatible elements, but this does not change the isotopic composition of the rising dehydrated solid. According to the isotopic differences, to fit Bercovici and Karato's model, the deep mantle rises at a very slow speed of about 1 mm per year or even less. So it would take more than 400 million years to traverse the 400-km distance towards the melting region beneath ocean ridges. But upwelling rates in the immediate vicinity of mid-ocean ridges are significantly greater, and this would tend to minimize the time available for isotopic differences to develop. Still, opinions differ rather widely as to how great the isotopic difference actually is between average mantle samples from mid-ocean-ridge basalts and from plumes. Moreover, it is likely that plume-source material does not represent average lower mantle; instead, its composition is probably biased towards segregated subducted former oceanic crust^{10,11}.

Another concern is that the model depends on the assumption that the mantle transition zone contains enough water to cause the postulated melting. There is at least circumstantial evidence that the mantle is rather severely dehydrated during subduction¹², and estimates based on trace-element ratios¹³ limit the water content of the primitive silicate mantle to less than about 0.05% — much lower than Bercovici and Karato's estimate of 0.1–1.5% for the transition zone. But these measurements may not necessarily be applicable to the transition zone in the new model.

Bercovici and Karato's hypothesis is intriguing, but wouldn't such a mid-mantle melt layer have been detected long ago by seismologists? Perhaps it was^{14,15}, but the evidence is patchy. Features less than 10 km thick are easily missed in seismic records and will have to be carefully searched for in the future.

Albrecht W. Hofmann is at the Max-Planck-Institut für Chemie, Postfach 3060, 55020 Mainz, Germany.

e-mail: hofmann@mpch-mainz.mpg.de

1. Bercovici, D. & Karato, S. *Nature* **425**, 39–44 (2003).
2. Graham, D. W. *Rev. Mineral. Geochem.* **47**, 247–317 (2002).
3. Montelli, R. et al. *Geophys. J. Int.* (in the press).
4. Hofmann, A. W. *Nature* **385**, 219–229 (1997).
5. Davaille, A. *Nature* **402**, 756–760 (1999).
6. Kellogg, L. H. et al. *Science* **283**, 1881–1884 (1999).
7. Tolstikhin, I. N. & Hofmann, A. W. *Geochim. Cosmochim. Acta* **66**, A779 (2002).
8. Morgan, J. P. & Morgan, W. J. *Earth Planet. Sci. Lett.* **170**, 215–239 (1999).
9. Williams, Q. & Hemley, R. J. *Annu. Rev. Earth Planet. Sci.* **29**, 365–418 (2001).
10. Hofmann, A. W. & White, W. M. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
11. Christensen, U. R. & Hofmann, A. W. *J. Geophys. Res.* **99**, 19867–19884 (1994).
12. Dixon, J. E. et al. *Nature* **420**, 385–389 (2002).
13. Michael, P. J. *Earth Planet. Sci. Lett.* **131**, 301–320 (1995).
14. Revenaugh, J. & Sipkin, S. *Nature* **369**, 474–476 (1994).
15. Vinnik, L. & Farra, V. *Geophys. Res. Lett.* **29**, doi: 10.1029/2001GL014064 (2002).

Cell biology

Join the crowd

R. John Ellis and Allen P. Minton

Cells are packed with large molecules. The ramifications of this 'crowding' for a wide range of intracellular processes are only now becoming more generally understood.

All cells contain lots of big molecules, especially proteins, nucleic acids and complex sugars. The very high total concentration of these molecules, or 'macromolecular crowding', has energetic consequences that could affect many aspects of cellular function. Yet most biochemical studies of macromolecular properties are carried out in dilute solutions in which crowding effects do not occur, despite the availability of polymeric compounds whose addition to such solutions mimics the phenomenon. The first international meeting* devoted to the biological implications of macromolecular crowding brought together 60 theoreticians and experimentalists; their interaction revealed both the diversity and the magnitude of crowding effects on cellular processes.

The high total concentration of macromolecules inside cells (up to 400 grams per litre) means that between 5% and 40% of the total volume is physically occupied by these molecules. An even larger fraction of the total volume is unavailable to other molecules of comparable size. For example, in a solution containing 30% by volume of identical globular molecules, less than 1% of the remaining volume is available to an additional molecule of equal size — that is, less than 1% can accommodate such a molecule without displacing one of the molecules already present. The work required to place the additional molecule in this solution is correspondingly much higher than that required to place it in a dilute solution.

However, the effect of volume occupancy on available volume is sensitive to the relative sizes and shapes of the occupying molecules. Any reactions that increase the available volume are theoretically stimulated by crowded conditions. These processes include the binding of macromolecules to one another, the folding of protein and nucleic-acid chains into more compact shapes, and the formation of aggregates, such as the amyloid deposits seen in some neurodegenerative diseases. Another effect of crowding should be to reduce the rate of diffusion by factors up to 10 — depending on the size of the diffusing particle and the degree of occupancy of the medium — compared with the rate in uncrowded buffers. Thus a biochemical reaction might be influenced by crowding



Figure 1 Crowded interior. This three-dimensional reconstruction shows part of the cytoplasm of an intact motile *Dictyostelium discoideum* cell. The orange linear complexes are actin filaments; ribosomes and other macromolecular complexes are in grey; membranes are in blue. Reprinted with permission from ref. 3.

if its rate is limited by diffusion or by the stability of macromolecular complexes. The magnitude of these various effects can be estimated by using equations developed to describe model fluids containing hard particles or randomly coiled polymers (A. Minton; J. Herzfeld, Brandeis Univ., Waltham, Massachusetts; R. de Vries, Wageningen Univ.).

How does reality conform to these theoretical expectations? Direct evidence for the crowded state of cell interiors is provided by the technique of cryoelectron tomography, in which thin intact cells are frozen rapidly in liquid ethane. The frozen cells are studied in an electron microscope that takes many pictures over a range of tilt angles; a computer program then reconstructs three-dimensional images with a resolution of 5–6 nanometres. The high density of actin filaments (part of the cellular 'skeleton') and ribosomes (protein-making machines) seen in reconstructed images of the slime mould *Dictyostelium* supports the view that the cytoplasm is filled with large ensembles of macromolecules, which form functional complexes, rather than with freely diffusing

and colliding macromolecules (S. Nickell, Max Planck Institute, Martinsried; Fig. 1). In addition, direct observation of fluorescent proteins in animal cells — both within the cytoplasm and inside two cellular compartments, the mitochondrion and endoplasmic reticulum — shows that their diffusion rate is reduced by factors in the range 3–8, broadly consistent with predictions (A. Verkman, Univ. California, San Francisco).

Higher levels of resolution for intact cells can be achieved with the use of nuclear magnetic resonance techniques to study the conformation of proteins expressed at high levels. For instance, a bacterial protein that lacks ordered structure in uncrowded buffers seems to acquire a degree of persistent structure when crowding agents are added to dilute solutions of the pure protein — and, more significantly, when expressed in the bacterium *Escherichia coli* (J. Bryant, Univ. North Carolina, Chapel Hill).

A type of crowding called confinement refers to situations in which macromolecules find themselves inside small compartments. Such compartments include those created by cytoskeletal structures or by the central cage of the chaperonin proteins — inside which newly synthesized proteins can fold, protected from crowding-enhanced, non-productive aggregation with other folding chains. Theory predicts that such confinement will stabilize compact shapes more than extended shapes and will enhance the rates of reactions leading to compaction (H.-X. Zhou, Florida State Univ., Tallahassee). These predictions agree qualitatively with observations that encapsulating proteins inside small pores in hydrated silica glasses enhances their stability when heated (D. Eggers, San José State Univ.; J.-M. Yuan, Drexel Univ., Philadelphia).

Another observation is that the efficiency of the bacterial chaperonins GroEL and GroES — which work together to help certain proteins fold — is enhanced by crowding agents. This might be due to an enhancement of the association between GroEL and GroES; the latter caps the internal cavity of GroEL and prevents the escape of an encapsulated polypeptide (J. Martin, Max Planck Institute, Tübingen). And it might be because crowding reduces the probability that an encapsulated polypeptide can diffuse away from an uncapped GroEL before folding completely (A. Elcock, Univ. Iowa, Iowa City).

* EMBO Workshop: Biological Implications of Macromolecular Crowding. Palacio de Magalia, Spain, 14–18 June 2003 (www.cib.csic.es/~revers/embo2003).

Crowding also seems to affect the way in which bacteria adapt to large changes in the concentration of osmotically active molecules in their environment: bacteria grown under widely disparate osmotic environments show significant differences in their intracellular concentrations of macromolecules, as well as of smaller molecules (S. Cayley, Univ. Wisconsin–Madison). Other processes influenced by crowding include: the regulation of metabolic pathways associated with signal transduction (H. Westerhoff, Free Univ. Amsterdam); the extraordinary stability of the crystallin proteins in the lens of the eye (J. Clauwaert, Univ. Antwerp); and the synthesis of compact proteins from peptide fragments by enzymes that catalyse the opposite reaction, polypeptide breakdown, in the absence of crowding (R. Roy, Natl Inst. Immunol., New Delhi). Moreover, when DNA is in the crowding-induced compact conformation — rather than the worm-like coiled shape seen in dilute solution — several enzyme-catalysed DNA-processing reactions are accelerated by many orders of magnitude (J. L. Sikorav, CEA-Saclay, Gif-sur-Yvette).

One of the more dramatic effects of crowding is the stimulation of the rate and extent of formation of rod-like protein aggregates. Examples are the amyloid fibres of synuclein that are implicated in Parkinson's disease (A. Fink, Univ. California, Santa Cruz), microtubules (D. Hall, Univ. Cam-

bridge), and large arrays of the protein FtsZ, essential to bacterial division (J. Gonzalez, CIB-CSIC, Madrid). Also, the rate of formation of fibres of deoxygenated haemoglobin S (the mutant form found in patients with sickle-cell anaemia) can vary over several orders of magnitude depending on the degree of crowding. This dependence can be accounted for quantitatively over the entire range of experimental observation by relatively simple hard-particle models of excluded volume (F. Ferrone, Drexel Univ.).

It seems likely from the observations reported at this meeting, and reviewed recently^{1,2}, that macromolecular crowding *in vivo* is involved in many aspects of cellular function. Given that crowding was not discovered yesterday, one may wonder why little mention of this phenomenon is found in current textbooks of biochemistry and molecular and cell biology. ■

R. John Ellis is in the Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

e-mail: jellis@bio.warwick.ac.uk

Allen P. Minton is at the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland 20892-0830, USA.

e-mail: minton@helix.nih.gov

1. Ellis, R. J. *Trends Biochem. Sci.* **26**, 597–604 (2001).

2. Hall, D. & Minton, A. P. *Biochim. Biophys. Acta* **1649**, 127–139 (2003).

3. Medalia, O. *et al. Science* **298**, 1209–1213 (2002).

expect. Also, as the temperature increases, the magnets heat up and their behaviour becomes more random, less correlated^{2,3}. We would therefore expect the susceptibility to decrease with temperature. Experiment shows that it does (Fig. 1a), but how do theoretical predictions compare?

One way of deriving the macroscopic properties of a physical system — known as the 'royal route' in the technical jargon — is by classical statistical mechanics. Here the first and most important task is to construct the partition function of the system under investigation. This is a sum over the various energy states that the system can occupy and tells us how the system distributes itself over the available states in terms of probabilities. Once we have the partition function, we can then compute all the other relevant (and macroscopically observable) properties, such as the internal energy of the system, its pressure, entropy and susceptibility and so on. This is quite remarkable: knowing the different energy states that the system can occupy is sufficient to construct all the other macroscopic quantities needed to describe the physical system completely. But, ultimately, this conclusion is erroneous, as can clearly be seen in the disagreement shown by Ghosh *et al.*¹ between the classical predictions and experimental results (Fig. 1a).

Quantum mechanically, if we wish to fully specify the state of a system, we need to know both the energy levels and the particular states corresponding to these energy levels. Loosely speaking, this means that for a salt such as $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$ we need to know the orientations of the magnetic atoms and not just their overall energy. More significantly, these states could, through quantum mechanics, display an excess of correlations between the individual magnets over and above anything allowed in classical physics — they could be entangled^{2,3}. And, amazing though it may seem, the presence of this entanglement can make a difference in observed macroscopic quantities. As the quantum correlations are stronger than classical ones, we can predict that the susceptibility will be higher according to quantum mechanics. This is indeed demonstrated by Ghosh *et al.*¹, and they also show that other quantities, such as the heat capacity, depend on the presence of entanglement (see Fig. 2 on page 49).

This work is important for at least two reasons: one is that it is no longer enough for physicists to investigate only the energy spectrum of a system, but some other features — such as entanglement in this case — are of paramount importance in the overall behaviour of the system. The other reason is that even a very small amount of entanglement can produce significant effects in the macroscopic world (Fig. 1b).

Finally, I should like to indulge in a little speculation. It is widely accepted that quantum mechanics is our most accurate

Quantum physics

Entanglement hits the big time

Vlatko Vedral

Entanglement is a quantum phenomenon usually associated with the microscopic world. Now it is clear that its effects are also relevant on macroscopic scales, such as in the magnetic properties of some solids.

Entanglement describes a correlation between quantum mechanical systems, such as photons or atoms, that does not occur in classical, newtonian physics. Under scrutiny since the birth of quantum theory, such correlations have been used to highlight a number of apparent paradoxes at the heart of quantum physics, but their existence has nevertheless been confirmed in a number of different experiments since the beginning of the 1980s. For instance, the entanglement of two photons means that once the state of one photon is known, we immediately know the state of the other; entanglement is also the basis of the quantum computer. But who, apart from possibly a few philosophers of physics and some computer scientists, actually cares about statistical correlations between two or more systems? In other words, do quantum correlations

affect anything of any importance in the real, macroscopic world? Amazingly, they do, and the paper by Ghosh *et al.*¹ on page 48 of this issue demonstrates just that.

Ghosh *et al.* report experiments involving the magnetic salt compound $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$ (see refs 4–6 in ref. 1). The atoms in this salt all behave like small magnets, interacting with each other as well as adjusting themselves to any external magnetic field (modelled by an Ising chain of interacting spins). The authors' investigated the magnetic susceptibility of the system for a range of low temperatures. Susceptibility tells us how much the magnets align with each other when an external magnetic field is applied, so the greater the susceptibility the more the magnets align when the external magnetic field is increased. Our intuition would tell us that the more correlated the magnets, the higher the degree of susceptibility we should

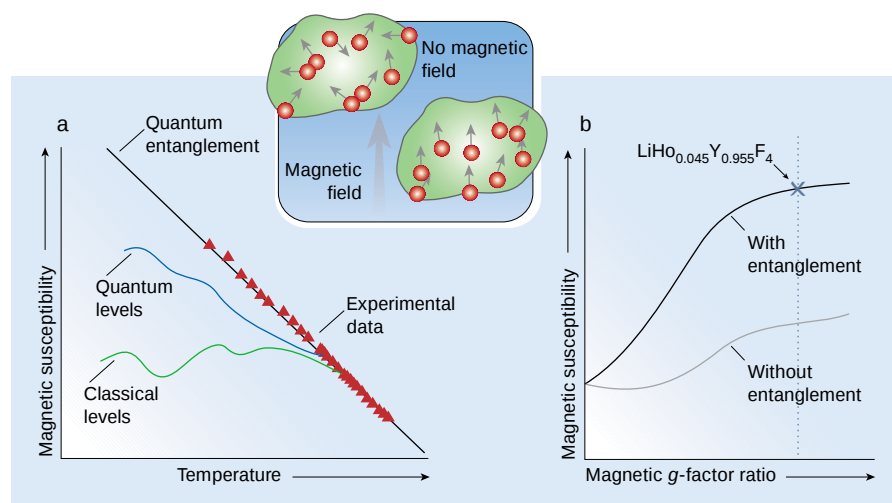


Figure 1 Far-reaching effects of entanglement. In the absence of a magnetic field, the magnetic moments of atoms in a solid point in random directions (inset); as a magnetic field is applied, the magnetic vectors tend to align with the field. But the alignment is greater than predicted by classical calculations, the excess correlation of the vectors being explained by the existence of quantum entanglement between them. Working with the salt $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$, Ghosh *et al.*¹ have demonstrated the influence of quantum entanglement on the macroscopic properties of a solid. a, Their measurements of the salt's magnetic susceptibility as a function of temperature are consistent with the prediction that includes quantum entanglement: models based on classical energy levels in the system, and on quantum energy levels but without entanglement, do not describe the data. b, The influence of entanglement on magnetic susceptibility is also seen through the ratio of magnetic 'g-factors' (which define the relationship between magnetization and spin in directions perpendicular and parallel to the applied magnetic field). The measured ratio and susceptibility (at fixed temperature) for $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$ are again consistent with the presence of entanglement. (Graphs derived from Figs 1 and 4 on pages 48 and 50.)

description of how atoms combine to form molecules, and it lies behind all chemistry. Chemistry in turn provides a basis for biological processes, including the metabolic cycles and the replication machinery making it possible for life to be sustained. So, might it be not only that quantum effects are responsible for the behaviour of inanimate matter, but that the magic of entanglement is also crucial in the existence of life⁴?

Vlatko Vedral is in the Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ, UK.
e-mail: v.vedral@ic.ac.uk

1. Ghosh, S., Rosenbaum, T. F., Aeppli, G. & Coppersmith, S. N. *Nature* **425**, 48–51 (2003).
2. Arnesen, M. C., Bose, S. & Vedral, V. *Phys. Rev. Lett.* **87**, 017901 (2001).
3. Gunlycke, D., Kendon, V., Vedral, V. & Bose, S. *Phys. Rev. A* **64**, 042302 (2001).
4. Landsberg, P. T. *Nature* **203**, 928–930 (1964).

Genetics

A balancing act

Janet Rossant

Mutations in model organisms are grist to the geneticists' mill: they help in assigning function to genes. Chromosome engineering in mice makes it easier to pinpoint the location of randomly induced mutations.

High throughput and high tech are the watchwords in the rush to develop new tools for genomic analyses. But traditional genetic screens of whole organisms still have a part to play. On page 81 of this issue, for instance, Kile and colleagues¹ show how tricks borrowed from fruitfly geneticists allowed the efficient isolation of many new mutations in mice.

Geneticists are always eager to have new mutations in 'model' experimental organisms to study, because mutations allow them to investigate how an organism's genetic make-up (its genotype) relates to its physical

and behavioural traits (its phenotype). One broad type of genetic screen is the phenotype-driven screen, in which mutations are generated at random across the genome, and offspring are scrutinized for phenotypes of interest. Mutant phenotypes that are heritable can then be mapped to the mutated region of the genome; next, the region is narrowed down, and the precise gene that is affected can be pinpointed.

This kind of approach has driven bacterial, yeast and invertebrate genetics for years, and has recently been applied to genome-wide screens in zebrafish. In mice,

the chemical *N*-ethyl-*N*-nitrosourea (ENU) is the mutating agent (mutagen) of choice for such screens; there are several large-scale ENU-based screens under way worldwide^{2,3}. Unexpected phenotypes, particularly relating to human disease, arise from ENU screens, because all kinds of mutations are generated, not just those that inactivate genes. For this reason, ENU-mutagenized eggs or sperm⁴ and embryonic stem cells^{5,6} can also be used in a sequence-based search for new variants (alleles) of known genes.

Phenotype-driven screens for dominant mutations (those in which just one of the two copies of a given gene need be mutated to produce a physical effect) can be performed on a large scale in mice, because phenotypes can be identified in the first-generation offspring of a mutated animal. Nevertheless, only about 1% of offspring reveal any sort of mutant phenotype, and good quantitative screening tools are needed to detect what can be fairly subtle effects.

By contrast, screens for recessive mutations, in which both copies of the gene need to be mutated to produce an effect, yield mutant phenotypes at a high rate. Such screens are an effective way of finding new mutations that cause embryonic death — even small screens have yielded mutations that identify new molecular players in development^{7,8}. But the drawback here is that it is hard to cover the whole genome, because of the numbers of mice involved and the difficulties in following pedigrees (family trees).

Tackling the genome one region at a time has advantages, as shown a few years ago⁹, in a strategy in which one part of a chromosome was deleted and mutations were detected on the other copy of that chromosome. Deletions of any size and position can now be engineered in the mouse genome. But many deletions turn out to be lethal when introduced as one copy into mice. It is thus not possible to use deletions to cover the whole genome for mutations. Instead, Kile *et al.*¹ took a leaf from the book of the fruitfly geneticists: they used a 'balancer' chromosome to enable them to isolate mutations in one particular chromosomal region easily, even though mutations were induced across the genome by ENU (Fig. 1, page 31).

By using chromosome-engineering strategies pioneered by Bradley and colleagues¹⁰, Kile *et al.* made an inversion within one of the two copies of their chromosome of interest, chromosome 11. The inversion spanned 2% of the genome, and an estimated 700 genes. It also disrupted the essential gene *Wnt3*, found at one end of the inverted sequence — so the inversion-bearing chromosome has to be balanced at all times by a normal copy of chromosome 11 (explaining why the inversion-bearing chromosome is called a balancer). Finally, the inversion was engineered to carry a foreign gene that results in yellow fur.

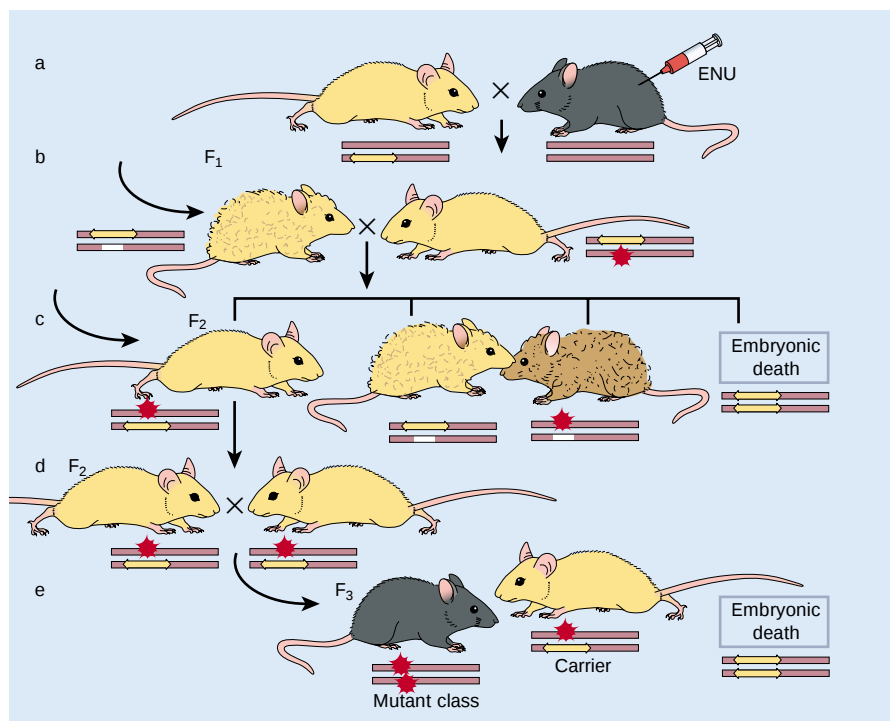


Figure 1 Using a balancer chromosome in mice: Kile and colleagues' strategy¹. **a**, Mice are generated in which one copy of chromosome 11 (purple) bears an inversion (yellow). The inversion inactivates the *Wnt3* gene, and carries a gene that produces yellow fur. These animals are crossed with mice in which mutations have been induced by the compound ENU. **b**, Some of the resulting (F_1) generation carry the balancer plus a mutated copy of the normal chromosome (right). These are mated with animals that bear the balancer plus a normal chromosome with a curly-coat gene (white, which marks the non-mutated normal chromosome). **c**, Some of the F_2 generation contain two balancer chromosomes, and die (because *Wnt3* is inactive). Some contain one balancer or one mutated chromosome plus the curly-coat chromosome. Others contain the mutated chromosome (all with exactly the same mutation) and balancer, and are identifiable by their yellow, non-curly coat. **d**, These are intercrossed. **e**, In the F_3 generation, animals with two identical mutant chromosomes are black, and are analysed to identify the defects caused by the mutation. Yellow mice can be bred to maintain the mutant line.

The authors crossed inversion-bearing females with ENU-mutagenized males. Then, any resulting yellow offspring (which inherited the balancer chromosome and a mutagenized 'normal' chromosome) were crossed with mice engineered to have one balancer chromosome plus one normal chromosome bearing a 'curly coat' gene. Any offspring of this cross that inherit two balancer chromosomes will die before birth (because *Wnt3* is disrupted). Surviving yellow mice with non-curly coats are chosen for further study, because they carry the balancer chromosome (from their parent bearing the curly-coat gene) and the mutagenized 'normal' chromosome (from their yellow-furred parent). When intercrossed, these mice have black offspring with two copies of the mutagenized chromosome. Lack of live black mice indicates an embryonic-lethal mutation. Yellow offspring can be retained as carriers.

This strategy vastly simplifies the breeding and maintenance of mutant strains — no complicated tools such as molecular genotyping are required, only the ability to follow coat colours. Similar balancer chromosomes are being generated by various groups for other regions of the mouse genome, and

these could be useful tools for maintaining any kind of mutant stock without genotyping, as is routinely done with fruitflies.

So, what mutations have been found so far in the region of interest? Kile *et al.*¹ report 88 new mutations from a screen of 735 pedigrees. Of these mutations, most (55) were lethal, 30 of these before birth. The other, viable, mutations fell into various classes, affecting for instance the skin, nervous system, blood cells, head and face, and fertility. Given that the phenotypic analysis of the surviving mice was by no means exhaustive, the proportion of lethal mutations may be an overestimate. However, comparison with known spontaneous and targeted mutations in the region suggests that it is not far out of line; the authors found 65 known mutations described in the literature (none that overlap with the mutations now identified): one-third are lethal, one-third produce phenotypically mutant but living animals, and one-third have no reported effects on phenotype.

With the availability of the mouse genome sequence, Kile *et al.* could also correlate one mutant phenotype, a type of anaemia, with mutations in a particular candidate gene — that encoding the solute carrier protein 4a1.



100 YEARS AGO

The Liverpool School of Tropical Medicine has just issued a most important and practical report upon the prevention of malaria in the tropics. Dr. Dutton, who conducted the expedition with conspicuous success, shows with striking clearness how a great deal of the disease is due to the want of knowledge of the nature of malaria, and that during the dry season the residents are largely to blame for the appearance of the disease... An account of the examination of one of the large compounds illustrates to what extent mosquitoes are bred by the white man in the tropics on his own premises... In one factory yard were found six barrels, and in the garden there were seventeen tubs and eight small wells, all breeding quantities of *Culex*, *Stegomyia* and *Anopheles* mosquitoes. Besides these dry season breeding places, discarded domestic utensils were scattered about the yard and garden which, in the wet season, would have acted as breeding places. It is pointed out that during the dry season, from November to May, natural breeding places for mosquitoes in Bathurst cease to exist and from this period the people breed mosquitoes solely in their own compounds. From *Nature* 3 September 1903.

50 YEARS AGO

An international Conference of Parapsychological Studies met in Utrecht during July 30–August 5 under the auspices of the University and with the financial support of the Parapsychology Foundation, Inc., of New York. Some sixty members of fourteen nationalities, comprising physicists, chemists, biologists, psychologists, philosophers, engineers and mathematicians took part in discussions of problems which arise in the investigation of those purported types of communication between individuals which are hitherto not explicable in terms of contemporary physics and biology... After preliminary sessions, four working groups were established: the first dealing with quantitative experimental studies of various classes of alleged paranormal activity; the second with the interpretation of material gathered in psychiatric practice; the third with qualitative and spontaneous (that is not experimentally controlled) phenomena; and the fourth with the psychological study of those persons who appear to display a large amount of paranormal sensitivity. From *Nature* 5 September 1953.

The use of the balancer chromosome makes it easier to find candidate genes such as this, because we know in what region they occur. Furthermore, the authors identified another 142 new ENU-generated mutations outside the inverted region. All of this represents a big boost to the mouse mutant resource.

The approach used by Kile *et al.* makes it easier to follow mutant genotypes in ENU-based screens. Yet, as with all genetic screens, the yield of useful mutations depends on the sensitivity of the tools for identifying mutant phenotypes. We are probably missing many subtle defects in adult mutant mice, whether ENU-induced or generated by targeting a specific gene — and we cannot afford to do so, given that these animals are supposed to be a key genetic model for learning about human disease. However, international efforts to develop improved screening tools are underway. For instance, the US National Institutes of Health sponsors several ENU mutagenesis centres, each of which focuses

on different but overlapping phenotypic screens. And the Eumorphia project, sponsored by the European Union, is developing physiological, pathological, behavioural, biochemical and imaging tools to reveal abnormalities in all organ systems. International cooperation in generating, analysing and archiving mutations will be crucial if mice are to remain the prime mammalian model for human disease. ■

Janet Rossant is at the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada.

e-mail: rossant@mshri.on.ca

1. Kile, B.T. *et al.* *Nature* **425**, 81–86 (2003).
2. Nolan, P. *et al.* *Nature Genet.* **25**, 440–443 (2000).
3. Hrabe de Angelis, M. *et al.* *Nature Genet.* **25**, 444–447 (2000).
4. Coghill, E. L. *et al.* *Nature Genet.* **30**, 255–256 (2002).
5. Chen, Y. *et al.* *Nature Genet.* **24**, 314–317 (2000).
6. Munroe, R. J. *et al.* *Nature Genet.* **24**, 318–321 (2000).
7. Eggenschwiler, J. T., Espinoza, E. & Anderson, K. V. *Nature* **412**, 194–198 (2001).
8. Liu, S. *et al.* *Development* **129**, 5839–5846 (2002).
9. Rinchik, E. M. & Carpenter, D. A. *Genetics* **152**, 373–383 (1999).
10. Zheng, B., Sage, M., Sheppard, E. A., Jurecic, V. & Bradley, A. *Mol. Cell. Biol.* **20**, 648–655 (2000).

such as the dynamics of air–sea interaction in the tropics and the sensitivity of the global system to changing greenhouse gases.

In 1999, a project known as EPILOG (Environmental Processes of the Ice Ages, Land, Ocean, Glaciers) began to reassess the state of the glacial world by using new and improved tools for telling time and monitoring temperature³. The programmatic seeds planted four years ago are now beginning to bear fruit. The GLAMAP-2000 project, which began independently but later joined EPILOG, focused on the Atlantic Ocean from Antarctica to the Arctic. Chronologies are greatly improved, based on higher-resolution records and better understanding of radiocarbon dating. Temperature estimates are refined, with updated versions of the statistical methods first developed by CLIMAP.

The GLAMAP-2000 results broadly confirm key CLIMAP findings, such as the relative stability of the subtropics and the impact of ice sheets on the North Atlantic, but they reveal significant summer melting of sea ice through the Norwegian Sea and into parts of the Arctic Ocean^{4,5}. Short-term variability within the glacial maximum was substantial, especially in the subpolar North Atlantic, where icebergs from ice-sheet margins occasionally inundated the ocean⁶.

New data suggest that sea ice was also highly seasonal in the Southern Ocean; although average winter ice advanced to about 50° S, about 5° north of its present limit, the position of the summer ice boundary was relatively stable⁷. Diatoms provide evidence for sporadic northward excursions of summer sea ice, and also for a maximum in average sea-ice extent that might have preceded the global glacial maximum by several thousand years.

For the lower latitudes, the GLAMAP-2000 data suggest that, compared with today, there was an extreme cooling in the eastern North Atlantic and along the eastern boundary upwelling system off Namibia, as well as substantial cooling of the South Equatorial Current in the central Atlantic^{8,9}. This spatial fingerprint of change echoes an independent reconstruction of the southeast Pacific, based on similar faunal methods, which found considerable ice-age cooling off Peru and along the Equator¹⁰. Although the average GLAMAP-2000 estimate of oceanic cooling (1.2 °C) is less than that of CLIMAP, the low-latitude cooling is much greater. The extent and timing of equatorial changes may be consistent with an idea that dynamics of the tropics trigger global climate change¹¹.

GLAMAP-2000 uses an 'inverse model'¹² to assimilate the sparse data on past temperatures into a simple model of the upper ocean. This pioneering effort uses wind parameters from an atmospheric general circulation model (AGCM) and assumes no change in deep-sea temperatures, a known limitation that must be addressed in the future.

Climate change

Chilled out in the ice-age Atlantic

Alan C. Mix

A new reconstruction of temperatures in the Atlantic Ocean during the last ice age, 21,000 years ago, is the latest act in providing climate modellers with details of the past to help predict the future.

Earth's ice ages reveal climate changes beyond the range of modern human experience, and provide insight into mechanisms of global change. Palaeoclimatologists have long sought an accurate map of sea surface temperatures at the height of the last ice age, 21,000 years ago, as this would provide a crucial test of the models used to simulate past, present and future climates. A collection of papers from the GLAMAP-2000 (Glacial Atlantic Ocean Mapping) project¹, published in *Paleoceanography*, presents the most detailed attempt yet to reconstruct temperatures for the Atlantic.

Producing such a map is quite a challenge, given the need to invent 'proxy' thermometers from the contents of sediments, ice cores and other traces of the past world. More than two decades ago, the celebrated CLIMAP project² created a first global map from the distribution of fossil plankton species. To estimate temperature from these data, CLIMAP developed a statistical calibration scheme known as a transfer function, modifications of which have been used ever since.

According to the CLIMAP results, the magnitude and even the direction of ice-age temperatures varied regionally. Evidence pointed to the largest temperature differences, compared with now, in the high North Atlantic region bordered by large ice sheets.

In the Greenland and Norwegian Seas, sea ice advanced farther south and persisted through the glacial summer. Seasonal 'melt-back' of sea ice was also diminished in the Southern Ocean. The centres of the subtropical gyres and the pools of warm water in the western reaches of the tropical oceans yielded the greatest surprise; on the evidence of CLIMAP's species-based methods, these areas were only a little cooler (and in some areas were even warmer) than in modern times.

Overall, CLIMAP found that, except at high latitudes, ice-age conditions were not so different from those of today, with global cooling averaging just 1.6 °C. The main inferred cause of the so-called 'polar amplification' of climate change was that the larger area of snow and ice produced a cooling feedback through increased reflection of solar energy.

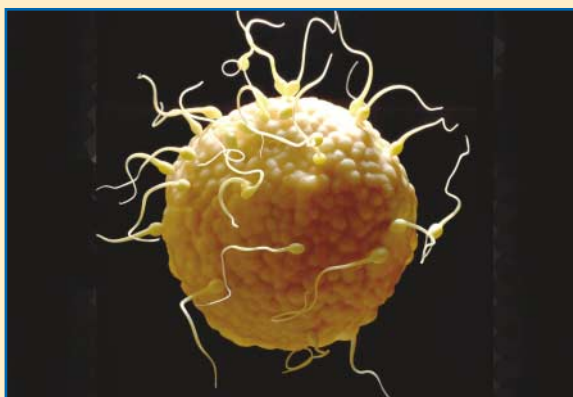
CLIMAP's reconstruction remains a benchmark boundary condition for climate models. Nevertheless, some researchers doubt that sea ice could have persisted year-round in the polar oceans. And modellers have been unable to reconcile CLIMAP's relatively small changes in tropical and subtropical sea surface temperature with independent evidence for greater cooling on land. These issues raise the profile of different proposed mechanisms for climate change,

Reproductive biology

Forming an attachment

Before a sperm can successfully fertilize an egg, it must bind to the egg's outer layer, the zona pellucida. This attachment is probably highly specific, as sperm will only recognize eggs from the same species. So what controls sperm–egg attachment? Writing in *Cell* (114, 405–417; 2003), Michael A. Ensslin and Barry D. Shur take a step towards answering this question — and their findings could offer clues to why some males with apparently 'normal' sperm are infertile.

In a hunt for egg-binding molecules, the authors looked at a protein from mouse testes, which they named SED1. This protein was a likely suspect because it was similar to a protein from boar sperm that had previously been shown to interact with components of the zona-pellucida. They found that



SED1 was located on the surface of mouse sperm — a good start. But more interesting was the finding that antibodies against SED1 blocked sperm–egg binding.

Of course, what happens in a test tube is not always important in animals, so the authors looked next at the effect of eliminating SED1 in

mice. They found that genetically engineered mice that lacked SED1 were much less fertile than normal mice. And although the sperm from these mice moved normally and were produced in the usual quantities, they failed to bind to eggs — proof that SED1 is indeed needed for sperm–egg binding. **Clare Thomas**

Although this approach does not provide a fully independent check on the data, it might eventually prove useful in diagnosing heat fluxes and other dynamic changes required to create such a pattern.

Additional models extend the GLAMAP-2000 results from the sea surface into the ocean interior. An AGCM that uses the new maps of sea surface temperature as a boundary condition drives an ocean general circulation model (OGCM) that produces a substantial cooling of water masses a few hundred metres below the sea surface, especially in the Southern Hemisphere¹³. Return of these chilly subpolar water masses to the sea surface at low latitudes provides a mechanism for cooling along the eastern boundary and the Equator, and points to possible interactions between low-latitude and high-latitude controls of climate. The OGCM also indicates that the deep sea was dominated by cold waters of Antarctic origin, although this finding depends on the assumed rate of sea-ice formation. Salt excluded from relatively fresh sea ice makes surface waters more dense, and thus more likely to sink.

A surprising result of the model is a poleward shift of the westerly winds and Antarctic Circumpolar Current during the ice age, in spite of known topographic steering of this current and an apparent expansion of cold water masses from the Southern Ocean towards the Equator. The inferred northward

extent of the cold Antarctic water masses suggests a possible deflection of Southern Ocean waters along the Pacific coast of South America, although this movement was not so large as to restrict the flow of relatively warm waters around South Africa⁷, an important conduit for global transports of heat and salt.

Recent experiments with coupled atmosphere–ocean models^{14–17} attempt to explain the glacial world while relying on temperature reconstructions only for verification of the model results. So far, the available coupled models produce divergent simulations of both the magnitude and direction of change in key features such as subtropical sea surface temperatures (ranging from slight warming¹⁴ to extensive cooling¹⁷) and deep-water formation in the North Atlantic (increased¹⁶ or decreased^{15,17}). Although the disagreements among the models, and between the models and data, underscore the need for caution in interpreting both 'hindcasts' of past climates and forecasts of future climates, they also point to the power of using a combined data-model approach to test, and improve, the models.

Finally, there is the question of how good the temperature data are. Conflicts occur in some areas between the temperatures reconstructed from fossil plankton species, and those from geochemical proxy measurements, such as the Mg/Ca ratio in foraminifera and an index, known as $U_{37}^{K'}$, based on traces of organic molecules

produced by certain plankton species. In the GLAMAP-2000 data set, discrepancies between the temperature proxies are especially acute for the Namibian upwelling system⁸ and the eastern equatorial Atlantic⁹. Similar discrepancies appear in some areas of the Pacific, leading to debate about whether the mean state of the glacial ocean was more akin to that of a modern interannual La Niña^{17,18} or El Niño^{19,20} event (cold or warm eastern equatorial region, respectively), or was something entirely different¹⁰.

The statistical basis of the transfer-function approach makes it best able to circumvent biases associated with seasonal plankton blooms. But in some cases the assumption that modern spatial variability serves to calibrate past biotic variations in terms of climate may break down. Culture experiments attest to the value of the geochemical tracers. But in the real world, organisms near their limits of thermal tolerance tend to adjust their depth or season of peak production, which would give the false impression of little or no temperature change in the chemical tracers.

The particular challenge for the coming years is to find a reliable means of understanding and correcting for the biases in each method of estimating temperature. With that, the resulting best estimate of large-scale climate changes will inform modelling experiments and reveal the primary processes responsible for global environmental change.

Alan C. Mix is at the College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331-5503, USA.
e-mail: amix@coas.oregonstate.edu

1. Sarnthein, M. *et al. Paleoceanography* published online 3 May 2003 (doi:10.1029/2002PA000769).
2. CLIMAP. *Geol. Soc. Am. Map Chart Ser.*, MC-36 (1981).
3. Mix, A. C., Bard, E. & Schneider, R. *Quat. Sci. Rev.* **20**, 627–657 (2001).
4. Nørgaard-Pedersen, N. *et al. Paleoceanography* published online 30 July 2003 (doi:10.1029/2002PA000781).
5. Sarnthein, M., Pflaumann, U. & Weinelt, M. *Paleoceanography* published online 11 June 2003 (doi:10.1029/2002PA000771).
6. Weinelt, M. *et al. Paleoceanography* (in the press).
7. Gersonde, R. *et al. Paleoceanography* published online 23 July 2003 (doi:10.1029/2002PA000809).
8. Niebler, H.-S. *et al. Paleoceanography* published online 27 August 2003 (doi:10.1029/2003PA000902).
9. Pflaumann, U. *et al. Paleoceanography* published online 2 August 2003 (doi:10.1029/2002PA000774).
10. Feldberg, M. & Mix, A. C. *Mar. Micropaleont.* **44**, 1–29 (2002).
11. Cane, M. *Science* **282**, 59–61 (1998).
12. Grieser, B. & Niebler, H.-S. *Paleoceanography* (in the press).
13. Paul, A. & Schäfer-Neth, C. *Paleoceanography* published online 9 July 2003 (doi:10.1029/2002PA000783).
14. Kitoh, A. & Murakami, S. *Paleoceanography* published online 11 September 2002 (doi:10.1029/2001PA000724).
15. Shin, S.-I. *et al. Clim. Dynam.* **20**, 127–151 (2003).
16. Hewitt, C. D., Stouffer, R. J., Broccoli, A. J., Mitchell, J. F. B. & Valdes, P. J. *Clim. Dynam.* **20**, 203–218 (2003).
17. Kim, S.-J., Flato, G. M. & Boer, G. J. *Clim. Dynam.* **20**, 635–661 (2003).
18. Martínez, I., Keigwin, L., Barrows, T., Yokoyama, Y. & Southon, J. *Paleoceanography* published online 14 May 2003 (doi:10.1029/2002PA000877).
19. Koutavas, A., Lynch-Stieglitz, J., Marchitto, T. & Sachs, J. *Science* **279**, 226–229 (2002).
20. Stott, L., Poulsen, C., Lund, S. & Thunell, R. *Science* **279**, 222–226 (2002).

Astrophysics

Planetary orbits do the time warp

Astrophys. J. (in the press); Preprint at <http://arXiv.org/astro-ph/0308112> (2003)

The orbits of Jupiter-like planets — or 'gas giants', which are known to exist in stellar systems other than our own — should often be tilted, instead of lying in a flat plane. So predict Edward W. Thommes and Jack J. Lissauer, using three-dimensional simulations of the motions of newly formed 'solar systems' of giant planets.

Systems of planets in orbit around a star feel resonant kicks as they are tugged repeatedly by each other's gravity, and the authors show that this can eventually cause the planets' orbits to misalign by tens of degrees relative to each other. The orbits are forced to change because the outermost giant planet gradually spirals in towards the central star, losing energy as it passes through the debris left over from the gas disk that spawned the system. As the outer planet migrates inwards, it influences the orbits of the inner planets, causing them to distort, synchronize and eventually oscillate far out of the system's original plane.

The authors predict that many 'exoplanet' systems around stars will turn out not to be coplanar, as this resonance-driven twisting of orbits should be widespread. **Joanne Baker**

Cancer

Location, location, location

Cancer Cell **4**, 133–146 (2003)

When tumours outgrow their oxygen supply they activate a 'hypoxic' response — controlled by the protein HIF — to stimulate new blood-vessel growth. Targeting HIF would seem to be a promising antitumour strategy. But according to Barbara Blouw *et al.*, the importance of this protein for tumour survival depends on where the tumour is located.

The authors looked at a class of malignant tumour called astrocytoma. The natural environment of this form of cancer is the brain, but many studies of astrocytoma development involve mouse models in which the tumours are grown under the skin. To investigate whether the hypoxic response

differs in these two environments, Blouw *et al.* implanted 'normal' cancerous astrocytes or cells genetically engineered to lack HIF-1 α under the skin or in the brains of mice.

Under the skin, the HIF-1 α -deficient tumours grew more slowly than the normal tumours. But the reverse occurred in the brain — here, loss of HIF-1 α caused the tumours to develop faster than usual and to invade more regions of the brain (see picture below). These tumours also possessed about 50% more blood vessels than the normal tumours.

So it seems that when studying the hypoxic response of tumours — like so much else in life — the choice of location is crucial. **Clare Thomas**

Immunology

Arresting asthma

J. Clin. Invest. **112**, 566–574 (2003)

Asthma is characterized by inflammation of the airways: it causes them to tighten, and airflow to be blocked. In people with allergic asthma, these events are triggered by allergens such as pollen or dust, which stimulate certain of the immune system's T cells to migrate to the lungs, proliferate, and release inflammatory chemicals.

T cells that lack the protein β -arrestin-2 show defects in migration *in vitro*, so Julia K. L. Walker *et al.* decided to study the protein's effects *in vivo*. They experimentally induced allergic asthma in normal mice and mice without β -arrestin-2, and found that healthy controls overreacted to the allergen and became asthmatic, whereas the mutant mice remained asthma-free — their lungs were devoid of T cells, and their airways were normal.

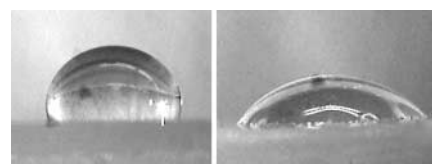
This suggests that the lack of β -arrestin-2 impairs the ability of T cells to migrate to the lungs, nipping inflammation in the bud and preventing mice from developing allergic asthma. The authors speculate that the protein could become an attractive drug target. **Helen R. Pilcher**

Surface science

What's cooking?

Langmuir doi:10.1021/la034138h (2003)

Once, scientists might only have used microwave ovens to heat up their lunch. But,



Better wetting the microwave way.

more and more, researchers are moving their kitchen appliances into the lab and coming up with inventive new uses for them. Brent T. Ginn and Oliver Steinbock have now shown that a microwave oven can generate a plasma that chemically and structurally modifies the surface of a polymer.

Polydimethylsiloxane (PDMS) is placed inside a pristine vacuum desiccator that contains a small piece of steel wire. The desiccator is purged with oxygen, evacuated and cooked at maximum power for up to 25 seconds. Sparks from the metal wire ignite the plasma. After this treatment, the PDMS surface is less rough (measured through profilometry) but more 'wetter' (measured using contact angles). The PDMS sample becomes less wettable over time, but there is a lasting improvement in its hydrophilicity.

The microwave technique is unlikely to deliver the kind of surface chemical control that can be achieved with more sophisticated plasma rigs. But this cheap and cheerful method of increasing surface wettability could be practical when only simple surface modification is required. **Rosamund Daw**

Evolutionary biology

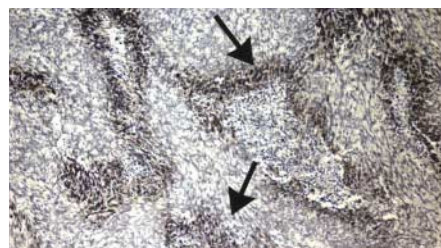
Loss for profit

Genetics **164**, 1271–1277 (2003)

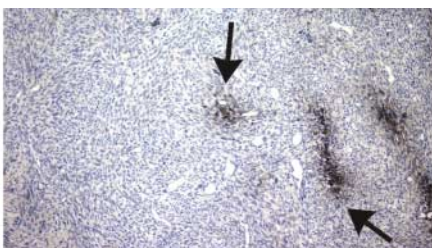
According to Erik R. Zinser and colleagues, natural selection sometimes favours the loss of beneficial genes. They have found that one such gene can be switched off when the genetic element that regulates it is expropriated by another gene.

Zinser and colleagues subjected cultures of the bacterium *Escherichia coli* to prolonged starvation. In one bacterium, a transposable element — a mobile stretch of genomic DNA — inserted itself into the regulatory region for a gene involved in processing peptides. This inactivated the gene, even though it is beneficial under starvation conditions. The gene's regulatory region then moved to another part of the genome, leading to another set of genes, useful for amino-acid consumption, being switched on — and allowing the mutant bacterium to prosper.

The researchers suggest that the mutant may have been aided by its initial rarity, allowing it to exploit a niche underused by the rest of the population. In this way, different bacterial strains might become specialized to take advantage of particular environmental conditions. **John Whitfield**



Normal (left) and HIF-1 α -deficient (right) tumours in the brain. The normal tumour has more oxygen-starved regions (arrows).



Horizontal prion transmission in mule deer

The gathering of deer during winter may foster the spread of chronic wasting disease.

Epidemics of contagious prion¹ diseases can be perpetuated by horizontal (animal to animal) and maternal (dam to offspring, before or after birth) transmission^{2–7}, but the relative importance of each mechanism is unclear. Here we compare the incidence of chronic wasting disease (CWD) in captive mule deer (*Odocoileus hemionus*) that is attributable to horizontal or maternal transmission. We find that horizontal transmission is remarkably efficient, producing a high incidence of disease (89%) in a cohort of deer in which maternal transmission was improbable. Our results indicate that horizontal transmission is likely to be important in sustaining CWD epidemics.

Although prion diseases have emerged worldwide as a threat to human and animal

health, they are incompletely understood. Of those recognized so far in mammalian species (including humans), only scrapie and CWD behave as contagious diseases; however, their control is impeded by a lack of basic knowledge about the transmission of the infective prion agent. Horizontal transmission apparently drives scrapie epidemics in sheep^{3,5,6}, but genetics also influence these patterns⁵. The epidemiology of CWD in mule deer resembles that of scrapie⁷, but the susceptibility of deer to the disease does not seem to be affected by genotype⁸.

It follows that CWD in mule deer (Fig. 1) represents a simple system for studying the mode of transmission of a contagious prion disease. We therefore compared the relative contribution of horizontal and maternal transmission of CWD in this species over a 5-year period (1997–2002). Captive mule deer in Colorado wildlife research facilities have been infected with CWD since the 1960s⁴. After attempts to eradicate CWD from the Foothills Wildlife Research Facility (FWRF) of the Colorado Division of Wildlife in 1985 (ref. 4), a new mule deer herd was started in 1990 with nine founders and augmented with fawn cohorts raised in 1991, 1992, 1993 and 1994. Despite preventive measures, CWD was diagnosed in 1994 in an FWRF-born female from the 1991 cohort. By May 1997, 10 cases had occurred among 57 resident deer (Fig. 2a).

We compared CWD incidence (defined as the number of animals affected divided by the number of animals at risk) in two sympatric mule deer cohorts. One cohort (R) of 10 deer was born to FWRF-resident dams in June 1997. Because the dams of all except one R-cohort deer eventually contracted CWD, maternal transmission was a possibility in this cohort.

The other cohort (NR) of 11 deer was also born in June, but away from the FWRF, and was added to the FWRF population in September–October 1997. Tonsil biopsies were negative in these NR deer, who were weanling fawns of about 3.5 months old that had been randomly captured from an enclosed, geographically separate herd in which CWD has not been detected (0 positive results out of 198 samples from deer aged one year or older, evaluated post mortem by immunohistochemistry of brain and lymphoid tissues; the 95% confidence interval for prevalence in this source herd was 0–1.9%). We presumed that dams of NR deer were not infected with CWD.

A comparison of these two cohorts shows that maternal transmission made little, if any, contribution to the occurrence of CWD.



Figure 1 Close quarters: proximity with fellow deer seems to be the overriding factor in prion transmission in these animals.

Despite the difference in potential for neonatal exposure to the disease, the overall CWD incidence was similar (Fisher's exact test, $P \approx 1.0$) in the R and NR cohorts: 9 out of 9 R individuals surviving for 1.8 years or longer developed CWD, as did 8 out of 9 NR individuals surviving for this length of time. The relative risk of CWD infection in R deer was 1.1 (95% confidence interval, 0.9–1.4).

Mean intervals to CWD-caused death did not differ ($z = -1.07$, $P = 0.28$) between the R and NR cohorts, but the dynamics of the epidemic differed slightly (Kaplan–Meier $\chi^2 = 3.74$, d.f. = 1, $P = 0.053$; Fig. 2b). This could be due partly to the differing ages at initial exposure for the two cohorts, an effect that is seen in scrapie^{3,5}.

Further evidence of horizontal transmission^{6,9} was provided by the increase in the cohort-specific incidence of CWD and the sharp decrease in the mean age of death from CWD in the FWRF herd (Fig. 2c).

The accumulation of abnormal prion protein in the peripheral lymphatic tissue of a host species seems to be strongly related to the horizontal transmissibility of prion diseases, and so might be a better predictor of contagiousness than the prion strain itself. The demonstration of abnormal prion protein in gut-associated lymphoid tissues^{10,11} but not in placental tissues¹¹ of mule deer infected with CWD is consistent with an alimentary shedding route⁸, as has been suggested for scrapie¹².

Both direct and indirect transmission of CWD can probably occur⁸, and concentrating deer in captivity or by feeding them artificially may facilitate transmission. The social behaviour of deer, particularly their tendency to gather and yard together during winter, might also foster epidemics in the wild. The efficiency of horizontal CWD transmission, exacerbated by behavioural ecology, might explain the high probability of CWD's becoming established in deer populations once it has been introduced either by natural spread or by the commercial movement of infected animals.

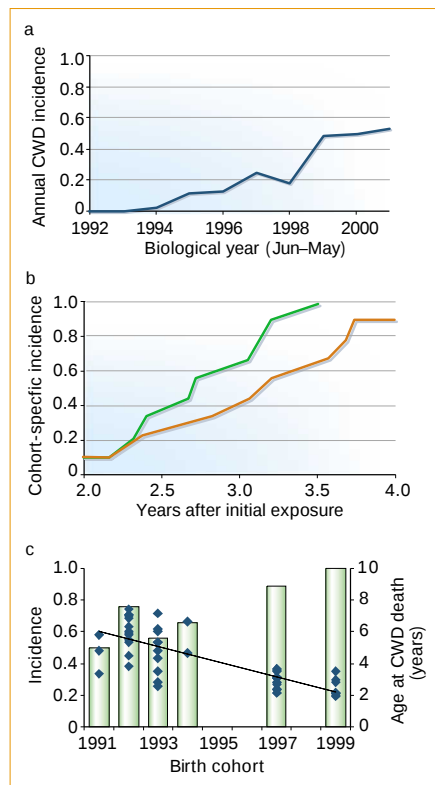


Figure 2 Horizontal transmission of chronic wasting disease (CWD) in captive mule deer. **a**, Annual incidence of CWD increased from 1994 to 2001. **b**, The cumulative incidence of CWD in deer in the R cohort (born to a CWD-infected dam; 100%; green line) and in the NR cohort (born to an uninfected dam; 89%; orange line) is similar, but the forms of the epidemic curves differ slightly. Deaths from CWD in R and NR deer occurred at 2.8 ± 0.2 years (range, 2.2–3.5 years) and 3.1 ± 0.2 years (range, 2.1–3.7 years) after exposure, respectively. **c**, Cohort-specific incidence (bars) increases and age at CWD-caused death declines (-0.46 yr^{-1} ; $P < 0.001$; diamonds) in 1991–99, which is consistent with horizontal transmission^{6,9}. No data are shown for 1995, 1996 or 1998 because fawns were not recruited in those years (c).

Michael W. Miller*, Elizabeth S. Williams†

*Colorado Division of Wildlife, Wildlife Research Center, Fort Collins, Colorado 80526, USA

e-mail: mike.miller@state.co.us

†Department of Veterinary Sciences, University of Wyoming, Laramie, Wyoming 82070, USA

1. Prusiner, S. B. *Proc. Natl Acad. Sci. USA* **95**, 13363–13383 (1998).

2. McFadyen, J. J. *Comp. Pathol. Ther.* **31**, 102–131 (1918).

3. Hourrigan, J., Klingsporn, A., Clark, W. W. & de Camp, M. in *Slow Transmissible Diseases of the Nervous System: Vol. 1* (eds Prusiner, S. B. & Hadlow, W. J.) 331–356 (Academic, New York, 1979).

4. Williams, E. S. & Young, S. *Rev. Sci. Tech. Off. Int. Epiz.* **11**, 551–567 (1992).

5. Hoinville, L. J. *Rev. Sci. Tech. Off. Int. Epiz.* **15**, 827–852 (1996).

6. Woolhouse, M. E. J., Stringer, S. M., Matthews, L., Hunter, N. & Anderson, R. M. *Proc. R. Soc. Lond. B* **265**, 1205–1210 (1998).

7. Miller, M. W. et al. *J. Wildl. Dis.* **36**, 676–690 (2000).

8. Williams, E. S. & Miller, M. W. *Rev. Sci. Tech. Off. Int. Epiz.* **21**, 305–316 (2002).

9. Redman, C. A. et al. *Epidemiol. Infect.* **128**, 513–521 (2002).

10. Sigurdson, C. J. et al. *J. Gen. Virol.* **80**, 2757–2764 (1999).

11. Spraker, T. R. et al. *Vet. Pathol.* **39**, 110–119 (2002).

12. Hadlow, W. J., Kennedy, R. C. & Race, R. E. *J. Infect. Dis.* **146**, 657–664 (1982).

Competing financial interests: declared none.

Nanotube electronics

Large-scale assembly of carbon nanotubes

Nanoscale electronic devices made from carbon nanotubes, such as transistors and sensors^{1–5}, are much smaller and more versatile than those that rely on conventional microelectronic chips, but their development for mass production has been thwarted by difficulties in aligning and integrating the millions of nanotubes required. Inspired by biomolecular self-assembly processes, we have created chemically functionalized patterns on a surface, to which pre-grown nanotubes in solution can align themselves in huge numbers. This method allows wafer-scale fabrication of millions of carbon-nanotube circuits with single-nanotube precision, and may enable nanotube-based devices, such as computer chips and high-density sensor arrays, to be produced industrially.

We used organic molecular marks on a substrate to guide the self-assembly of individual single-walled carbon nanotubes (swCNTs; see supplementary information). In the surface-functionalization step, we created two distinct surface regions coated either with polar chemical groups (such as amino (–NH₂/–NH₃⁺) or carboxyl (–COOH/–COO[–])) or with non-polar groups (such as methyl (–CH₃)). We achieved this by direct deposition of proper organic molecules (for example, as a self-assembled monolayer) by dip-pen nanolithography^{6–8} or by microcontact stamping⁹. These deposition techniques enabled us to functionalize the substrates without resorting to intermediate chemical steps, thereby minimizing surface contamination.

When the substrate is placed in a suspension of swCNTs, the nanotubes are attracted towards the polar regions and self-assemble to form pre-designed structures, usually within about 10 s. We used a magnet to remove common magnetic nanoparticle impurities from swCNT suspensions to improve the reliability of the process.

We discovered that a lateral-directional force exists on swCNTs near the boundary between the polar and non-polar molecular

regions (Fig. 1a). This force, which presumably originates from electrostatic interactions, rotates the swCNTs towards the polar region and confines them to the inside of it (Fig. 1a, inset).

Previous methods have relied on external forces, such as electric or magnetic fields, and liquid flow to align nanowires precisely^{10–13}. However, it is time-consuming to align millions of randomly oriented nanotube circuits by using external forces. In our process, individual polar molecular marks attract and align swCNTs along pre-determined lines without external force, enabling any swCNT-based structure to be assembled simply by using polar molecular patterns with the required shapes.

We scaled up this process for high-throughput assembly. In principle, large numbers of microscale molecular patterns can easily be generated over a large chip area by using high-throughput patterning methods such as photolithography, stamping and, in the future, parallel dip-pen nanolithography⁸.

Our assembly of millions of individual swCNTs on stamp-generated microscale patterns that cover areas of about 1 cm² on gold, for example, occurs with a yield of more than 90% (Fig. 1b). Surprisingly, we found that in swCNT suspensions at low concentrations (for example, about 0.02 mg ml^{–1} in

1,2-dichlorobenzene), only a single nanotube lies at the centre of each microscale polar molecular pattern, even though there is enough room for more nanotubes to assemble (Fig. 1b, inset). We used height profiles, determined by atomic-force microscopy, to confirm that assembly originated from one nanotube. Presumably, the hydrophobic surface of the swCNT passivates the polar pattern on the substrate and reduces the likelihood of adhesion by additional swCNTs.

We incorporated this process into conventional microfabrication methods to make millions of swCNT-based circuits over areas of about 1 cm². Individual swCNTs are assembled between two polar molecular patterns generated by stamping on microfabricated gold electrodes. We used short (about 1 nm) polar molecules with π -electrons (such as 2-mercaptoimidazole) to minimize the contact resistance between the nanotubes and the electrodes. Atomic-force microscopy, using a conducting probe, confirmed the existence of stable swCNT circuits that conducted micro-ampere currents.

Figure 1c shows that more than 70% of the junctions are connected by only one swCNT. This proportion should increase as swCNT-purification processes advance or as the gap size is reduced for high-density integration (for example, a roughly 90% yield on flat surfaces). High-precision assembly should then enable swCNTs to be used in practical applications.

Saleem G. Rao, Ling Huang,

Wahyu Setyawan, Seunghun Hong*

Department of Physics, Center for Materials Research and Technology and Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32306, USA

*Present address: Department of Physics, Seoul National University, Seoul 151-742, South Korea
e-mail: shong@phy.snu.ac.kr

1. Tans, S. J. et al. *Nature* **386**, 474–477 (1997).

2. Lee, J. et al. *Nature* **415**, 1005–1008 (2002).

3. Kong, J., Soh, H. T., Cassell, A. M., Quate, C. F. & Dai, H. *Nature*

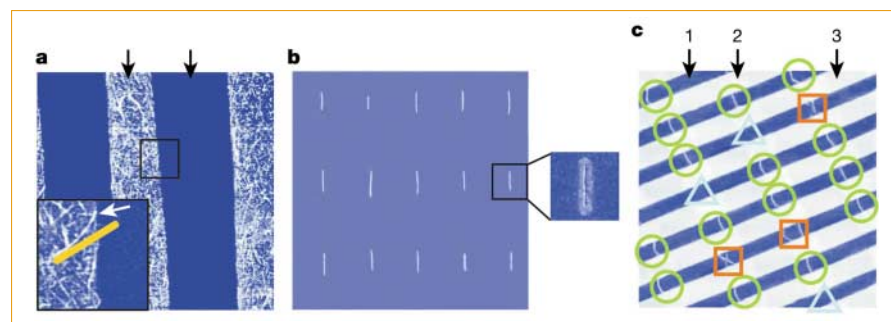


Figure 1 Atomic-force micrographs showing large-scale self-assembly of single-walled carbon nanotubes (swCNTs). **a**, Image (12 × 12 μm²) showing the topography of swCNTs near the boundary (white arrow, inset) between polar (cysteamine; left arrow) and non-polar (1-octadecanethiol (ODT); right arrow) molecular patterns on gold. No swCNTs are evident in ODT regions. Yellow bar represents a tangent to a bent swCNT, showing the extent of bending due to lateral-directional force. **b**, Topography (30 × 30 μm²) of an array of individual swCNTs covering about 1 cm² of gold surface. The friction-force image (inset) shows a single swCNT (dark line), and the regions containing 2-mercaptoimidazole (bright area) and ODT (dark area). **c**, Topography (20 × 20 μm²) of an array of junctions with no swCNTs (triangles), one swCNT (circles) or two swCNTs (squares), covering an area of about 1 cm². Arrows 1, 2 and 3 indicate octadecyltrichlorosilane (used to passivate the SiO₂ surface), 2-mercaptoimidazole on gold, and ODT on gold, respectively.

- 395, 878–881 (1998).
- Fuhrer, M. S. *et al. Science* **288**, 494–497 (2000).
 - Bockrath, M. *et al. Science* **291**, 283–285 (2001).
 - Hong, S. & Mirkkin, C. A. *Science* **288**, 1808–1811 (2000).
 - Manandhar, P. *et al. Phys. Rev. Lett.* **90**, 115505 (2003).
 - Zhang, M. *et al. Nanotechnology* **13**, 212–217 (2002).
 - Xia, Y. & Whitesides, G. M. *J. Am. Chem. Soc.* **117**, 3274–3275 (1995).
 - Liu, J. *et al. Chem. Phys. Lett.* **303**, 125–129 (1999).
 - Chen, X. Q., Saito, T., Yamada, H. & Matsushige, K. *Appl. Phys. Lett.* **78**, 3714–3716 (2001).
 - Hone, J. *et al. Appl. Phys. Lett.* **77**, 666–668 (2000).
 - Huang, Y., Duan, X. F., Wei, Q. Q. & Lieber, C. M. *Science* **291**, 630–633 (2001).
- Supplementary information** accompanies this communication on Nature's website.
- Competing financial interests:** declared none.

Innate defence

Evidence for memory in invertebrate immunity

Acquired immunity in vertebrates is characterized by immunological memory and specificity, whereas the innate defence systems of invertebrates are assumed to have no specific memory^{1–3}. Here we use a model system of a copepod, which is a minute crustacean, and a parasitic tapeworm to show that the success of reinfection depends on the antigenic resemblance between the consecutively encountered parasites. This finding indicates that an invertebrate defence system may be capable of specific memory.

The degree of specificity and memory in invertebrate immunity is unclear^{3–5}. Although a defence reaction can be induced in invertebrates against pathogens⁶, the responses may not be specific as they are able to distinguish only between different classes of pathogen^{2,3}. However, genotype-specific interactions between invertebrate hosts and different lines of parasites are possible^{4,7}. Such coevolutionary phenomena are difficult to explain unless there is some specificity in the immune system.

We investigated the line-specific memory of the defence system of an invertebrate host, the copepod *Macrocyclus albidus*, against a natural parasite, the tapeworm *Schistocephalus solidus*. To vary the antigenic features of the pathogen presented to the host, we exposed each copepod to three tapeworm larvae and then either to three sibling parasites or to unrelated parasites of a different sib-group three days later (Fig. 1a). All tapeworms used for the second exposure were fluorescently labelled to distinguish them from the parasites used for primary infection⁸.

If there is a specific memory inherent in the defence by the copepod hosts, we would expect a reduction in the success of reinfection by the sibling parasites. Indeed, we found that prior exposure to related parasites resulted in less secondary infection than occurred after exposure to unrelated para-

sites (Fig. 1b). On average, the reinfection success was reduced from $59.5 \pm 3.5\%$ to $47.6 \pm 4.8\%$ of copepods infected (paired *t*-test, $t = 3.236$, $n = 24$, $P = 0.0037$). In addition, the average intensity of reinfection decreased from 0.81 ± 0.06 to 0.66 ± 0.07 tapeworms per host ($t = 2.723$, $n = 24$, $P = 0.0121$). This effect should increase with the antigenic similarity between the consecutively encountered parasites.

Because the tapeworms are simultaneous hermaphrodites whose self-fertilization (selfing) leads to homogeneous offspring, we compared the size of the host's 'memory' effect towards tapeworms produced by selfing against that for tapeworms produced by outcrossing. As predicted, the reduction in reinfection increased for selfed worms (Fig. 1c; *t*-test, $n = 8$; $n = 16$ for outcrossed worms; reinfection success, $t = 1.361$, $P = 0.187$; intensity, $t = 2.230$, $P = 0.036$).

Could this reduced reinfection be due to factors other than the host defence system? Neither host mortality nor primary infection (determined in 16 of the 24 sibships)

differed between treatments (nominal logistic models, $P > 0.2$). Could the parasites themselves cause the reduction in reinfection? This is unlikely because cooperation between kin should facilitate rather than reduce reinfection.

If within-host competition between siblings is particularly strong, only those hosts that did not clear the primary infection would be affected. Excluding hosts with a resident primary infection from our sample did not diminish the effect: prior exposure to sib-group parasites still reduced reinfection from 64.1% to 45.6% in this smaller sample (in a nominal logistic model, including sibship, the effect of previous exposure was $\chi^2_1 = 7.728$, $n = 128$, $P = 0.0054$). We conclude that parasite-derived effects are unlikely to explain the observed reduction in reinfection after consecutive exposure of the host to related parasites.

Our results show that the defence system of copepods can react more efficiently after it has previously encountered antigenically similar parasites. To our knowledge, this

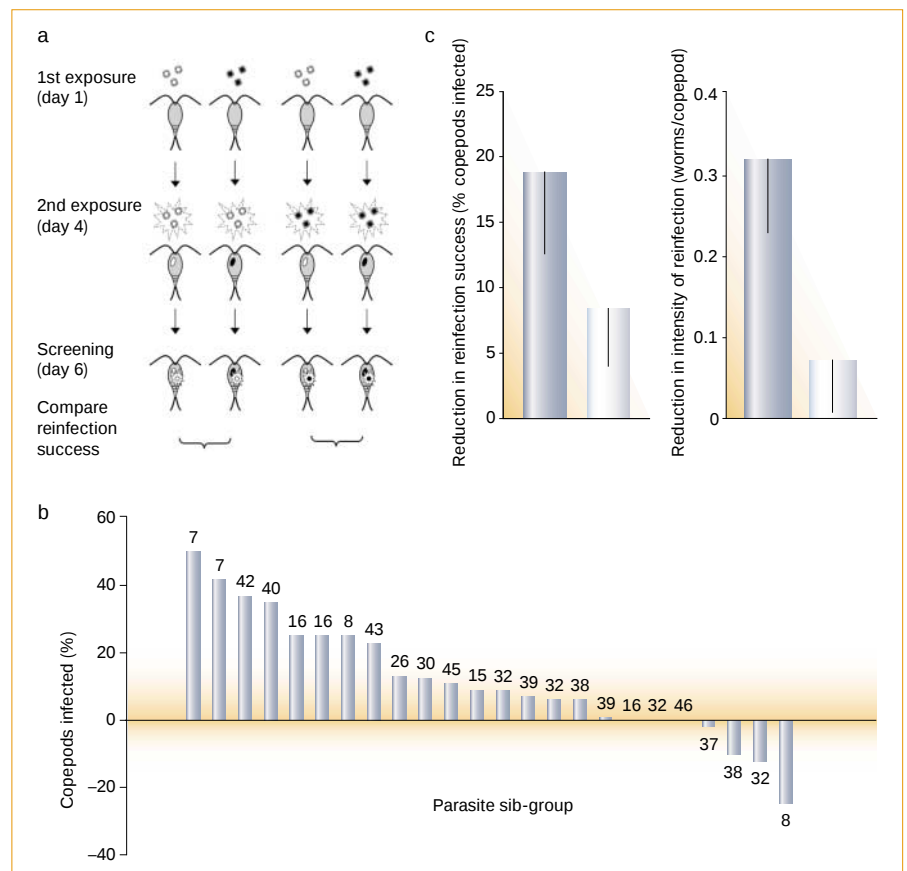


Figure 1 Exposure of a copepod to a tapeworm parasite reduces the chances of reinfection by antigenically similar tapeworm larvae. **a**, Experimental design. Two parasite sib-groups are shown (denoted by white and black larvae). There were $n = 24$ sib-groups, which varied in infectivity (nominal logistic model, $\chi^2_{23} = 95.95$, $n = 684$, $r^2 = 0.102$, $P < 0.001$). The experiment was reproduced in two rounds, both of which gave similar results and so were combined. Comparisons are made within the same parasite sib-group to control for variation in infectivity (further details are available from the authors). **b**, Reduction in reinfection due to prior exposure of the copepod host to sibling parasites, compared with exposure to unrelated tapeworms, for each of the 24 tapeworm sib-groups (arranged in rank order of difference). Sample sizes (number of copepods) are indicated above each bar. **c**, Difference in the success (left) and intensity (right) of copepod reinfection due to the degree of antigenic resemblance between tapeworm larvae used for reinfection and those used for primary infection. Left bars, selfed larvae; right bars, outcrossed larvae. Standard errors are indicated.

is the first demonstration of line-specific memory in individual invertebrates. In a study based on only two bacterial strains, maternally transferred immunity seemed to be strain-specific⁹. Invertebrates are occasionally capable of a specific response in the context of allogenic recognition¹⁰, but specificity in these cases is based on compatibility factors that have evolved within the species and so is not directly comparable with defence against a parasite.

The mechanism responsible for the line-specific memory in this copepod remains to be investigated, although binding by lectins is one possibility¹¹. If specific memory in invertebrate defence is more widespread than originally thought, it would have implications for our understanding of innate immunity, host–parasite coevolution, and invertebrate-vector diseases such as malaria.

Joachim Kurtz, Karoline Franz

Department of Evolutionary Ecology, Max Planck Institute of Limnology, August-Thienemann Strasse 2, 24306 Plön, Germany

email: kurtz@mpil-ploen.mpg.de

1. Janeway, C. A., Travers, P., Walport, M. & Capra, J. D. *Immunobiology: The Immune System in Health and Disease* (Current Biology, London, 1999).
2. Hoffmann, J. A. & Reichhart, J. M. *Nature Immunol.* **3**, 121–126 (2002).
3. Klein, J. *Scand. J. Immunol.* **29**, 499–505 (1989).
4. Schmid-Hempel, P. & Ebert, D. *Trends Ecol. Evol.* **18**, 27–32 (2003).
5. Rolff, J. & Siva-Jothy, M. T. *Science* **301**, 472–475 (2003).
6. Faulhaber, L. M. & Karp, R. D. *Immunology* **75**, 378–381 (1992).
7. Carius, H. J., Little, T. J. & Ebert, D. *Evolution* **55**, 1136–1145 (2001).
8. Kurtz, J., van der Veen, I. T. & Christen, M. *Exp. Parasitol.* **100**, 36–43 (2002).
9. Little, T. J., O'Connor, B., Colegrave, N., Watt, K. & Read, A. F. *Curr. Biol.* **13**, 489–492 (2003).
10. Scofield, V. L., Schlumpberger, J. M., West, L. A. & Weissman, I. L. *Nature* **295**, 499–502 (1982).
11. Marques, M. R. F. & Barracco, M. A. *Aquaculture* **191**, 23–44 (2000).

Competing financial interests: declared none.

Microfluidic systems

High radial acceleration in microvortices

Microfluidic systems can conveniently be used for rapid analysis of biological samples^{1–5}. Here we describe a single re-circulating flow, or microvortex, that can generate a maximum fluid rotational velocity of up to 12 m s^{−1} and a corresponding radial acceleration in excess of 10⁶g. Such microvortices may be exploited in centrifugal microdevices to investigate the effects of high radial acceleration on biological and chemical processes.

A microvortex was generated in a diamond-shaped microchamber (Fig. 1a), which was visualized by using fluorescent polystyrene beads (diameter, 2 μm; Fig. 1b, left). The average linear velocity of the pressure-driven flow in the 30-μm-wide channel was 3 m s^{−1}. The diamond-shaped

notch along the side wall of an otherwise smooth channel causes flow detachment at the opening of the notch.

Unlike macroscale lid-driven cavity flow⁶, the combination of extremely tight rotation radii ($r < 50 \mu\text{m}$) and high rotational velocity ($v > 10 \text{ m s}^{-1}$) of fluid flow from constricted microchannels can lead to generation of radial accelerations (v^2/r) as high as 10^7 m s^{-2} . As the average linear velocity in the 30-μm-wide channel was increased from 3 to 45 m s^{−1} (Reynolds number (Re) = 245 at the microchamber opening), the polystyrene beads (density, $\rho = 1.05 \text{ g cm}^{-3}$) in water were spun out of the microchamber owing to high centrifugal acceleration (Fig. 1b, right).

To illustrate the potential applications of this microvortex system, we flowed red polystyrene beads ($\rho = 1.05 \text{ g cm}^{-3}$) and green silica beads ($\rho, 1.8\text{--}2.0 \text{ g cm}^{-3}$) in a solution of 50 % (w/w) of CsCl ($\rho = 1.56 \text{ g cm}^{-3}$;

Fig. 1c). As the average flow velocity in the 30-μm-wide channel was increased from 1.5 m s^{−1} ($\text{Re} = 13$ at the cavity's opening) to 20 m s^{−1} ($\text{Re} = 122$ at the cavity's opening), the red fluorescent beads were concentrated towards the centre of the vortex, whereas the green beads were spun towards the chamber's outer edge. Although the linear flow velocity was 20 m s^{−1}, the corresponding volumetric flow rate was only about 18 μl s^{−1}, owing to the small cross-sectional area of the 30-μm-wide channel.

A simple calculation of the radial acceleration from the average flow rates in the microchannel overestimates the acceleration that we observed experimentally, mainly because of the damping of the flow velocity near the walls of the microchannels for pressure-driven flows. Because the flow that drives the vortex at the opening of the diamond-shaped cavity is near the 30-μm-wide channel wall, the maximum fluid rotational velocity inside the cavity in Fig. 1b, c is only about 20% of the average flow velocity in the channel.

To attain ultrahigh rotational velocities in the microvortex, we designed a trapezoidal microchamber connected to a constricted microchannel that was placed asymmetrically along the width of a 200-μm × 200-μm channel (see supplementary information). Using this design, we measured maximum rotational velocities as high as 12 m s^{−1} at a distance of 10 μm from the vortex core, which corresponds to a radial acceleration of $1.4 \times 10^7 \text{ m s}^{-2}$ or $1.4 \times 10^6 g$ (Fig. 1d). We measured rotation rates (Fig. 1d), both by the time taken for a photolysed (uncaged) plug of fluorescein dye to travel a calibrated distance (5 μm) from a focused nitrogen photolysis laser beam to an argon laser probe volume (Fig. 1d, squares), and by the time taken for the plug of dye to travel completely around the microvortex (Fig. 1d, open circles)⁷ (and see supplementary information).

The ability to use fluid flow with no moving parts to generate large centrifugal forces should prove useful for integrated micro-analytical separations — for example, in sample preparation.

J. Patrick Shelby, David S.W. Lim, Jason S. Kuo, Daniel T. Chiu

Department of Chemistry, University of Washington, Seattle, Washington 98195-1700, USA

e-mail: chiu@chem.washington.edu

1. Reyes, D. R., Iossifidis, D., Auroux, P.-A. & Manz, A. *Anal. Chem.* **74**, 2623–2636 (2002).
2. Auroux, P.-A., Iossifidis, D., Reyes, D. R. & Manz, A. *Anal. Chem.* **74**, 2637–2652 (2002).
3. Thorsen, T., Maerkl, S. J. & Quake, S. R. *Science* **298**, 580–584 (2002).
4. Beebe, D. J., Mensing, G. A. & Walker, G. M. *Annu. Rev. Biomed. Eng.* **4**, 261–286 (2002).
5. Ng, J. M. K., Gitlin, I., Stroock, A. D. & Whitesides, G. M. *Electrophoresis* **23**, 3461–3473 (2002).
6. Shankar, P. N. & Deshpande, M. D. *Annu. Rev. Fluid Mech.* **32**, 93–136 (2000).
7. Shelby, J. P. & Chiu, D. T. *Anal. Chem.* **75**, 1387–1392 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

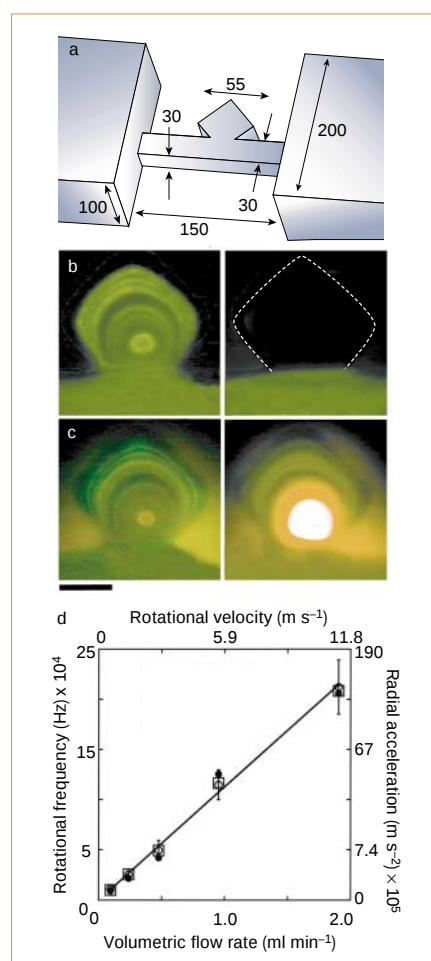


Figure 1 Generation and application of microvortices. **a**, Microchannel used to create the microvortex shown in **b**, **c**; dimensions are in μm. **b**, Fluorescent beads are spun out of the chamber as the flow velocity increases (from left to right). **c**, Separation of low-density red beads from high-density green beads in the microvortex as the flow velocity increases (from left to right). Scale bar, 25 μm. **d**, The experimentally measured rotational rates (squares and open circles) show excellent agreement with the vorticity results (filled circles) extracted from our simulations (some symbols overlap).

Whole-mantle convection and the transition-zone water filter

David Bercovici & Shun-ichiro Karato

Department of Geology and Geophysics, Yale University, PO Box 208109, New Haven, Connecticut 06520-8109, USA

Because of their distinct chemical signatures, ocean-island and mid-ocean-ridge basalts are traditionally inferred to arise from separate, isolated reservoirs in the Earth's mantle. Such mantle reservoir models, however, typically satisfy geochemical constraints, but not geophysical observations. Here we propose an alternative hypothesis that, rather than being divided into isolated reservoirs, the mantle is filtered at the 410-km-deep discontinuity. We propose that, as the ascending ambient mantle (forced up by the downward flux of subducting slabs) rises out of the high-water-solubility transition zone (between the 660 km and 410 km discontinuities) into the low-solubility upper mantle above 410 km, it undergoes dehydration-induced partial melting that filters out incompatible elements. The filtered, dry and depleted solid phase continues to rise to become the source material for mid-ocean-ridge basalts. The wet, enriched melt residue may be denser than the surrounding solid and accordingly trapped at the 410 km boundary until slab entrainment returns it to the deeper mantle. The filter could be suppressed for both mantle plumes (which therefore generate wetter and more enriched ocean-island basalts) as well as the hotter Archaean mantle (thereby allowing for early production of enriched continental crust). We propose that the transition-zone water-filter model can explain many geochemical observations while avoiding the major pitfalls of invoking isolated mantle reservoirs.

The two main sites of mantle upwelling in the Earth are mid-ocean ridges, such as the East Pacific Rise, which draw mantle materials up from shallow depths, and intraplate hotspots, or ocean islands, such as Hawaii, which are possibly due to deep-seated, upwelling plumes. The lavas sampled from these two environments, mid-ocean-ridge basalts (MORB) and ocean-island basalts (OIB), respectively, are in general chemically distinct; that is, MORBs are depleted in incompatible elements including uranium and thorium, while OIBs are relatively enriched in these elements. This suggests that MORB and OIB originate from different source regions.

The terrestrial heat flow also suggests the presence of different source regions. In particular, the approximately 36 TW of heat output from the mantle cannot be produced by a mantle entirely composed of MORB source materials, because it has so few radioactive heat sources; the remaining heat sources are thus thought to be hidden at greater depths in isolated reservoirs unsampled by mid-ocean ridges, but possibly sampled by plumes. In total, these and other observations suggest that the mantle is separated into at least two chemically distinct reservoirs, one depleted by melting processes at ridges, the other relatively enriched and untapped.

Although for many years the 660-km-deep discontinuity was believed to be the barrier between these reservoirs, recent tomographic studies suggested that subducting slabs cross this boundary^{1,2} and sink possibly as far as the core-mantle boundary³. Other mechanisms for retaining distinct reservoirs have been proposed, including poor mixing (the “plum pudding” model⁴) and placing the boundary between reservoirs near the bottom of the mantle at the D'' layer^{5,6}. One of the most recent mantle-layering models⁷ places the barrier between reservoirs at about 1,600 km depth; the advantages of this model as opposed to others are discussed in ref. 7. However, whereas the 1,600-km barrier is inferred by the depth at which slabs lose much of their tomographic signature (and are thus presumably stopped) the boundary is not observed in global seismic velocity models^{8,9} against which tomographic models are treated as perturbations. This lack of evidence has been attributed to large undulations in the boundary (due to impinging slabs and convective “lava-lamp” type motion¹⁰), causing an effectively broad, diffuse and hence undetectable boundary. Lack of observation of

these undulations in tomography are attributed to the somewhat fortuitous cancellation of chemical and thermal effects in causing seismic velocity anomalies¹¹.

Layered-mantle models also suffer from dynamical inconsistencies, regardless of whether the boundary between the layers is at 660 km, 1,600 km, or 2,700 km depth. Perhaps most importantly, because any enriched lower layer has most of the heat-producing elements, the depleted overlying layer is heated almost entirely along its base¹¹. Basally heated convective systems are characterized by upwelling and downwelling currents with comparable heat transport. However, various analyses of mantle plume heat flow suggest that they account for less than 10% of the net heat flux out of the mantle¹², the remainder being mostly transported by slabs. In general, evidence for slab-dominated convective circulation is very difficult to reconcile with layered-convection models that have a depleted layer basally heated by an underlying enriched one¹¹.

Although the conflict between geochemical and geophysical constraints has engendered great activity in the mantle dynamics community, the problem remains unresolved, and to some extent—given the ongoing appeal to mantle layering—is no further along than 20 years ago. We propose an alternative hypothesis that avoids the requirement for distinct chemical reservoirs and mantle layering, and explains the observations through a mechanism we term the “transition-zone water filter” (Fig. 1).

Water in the transition zone

The mantle transition zone is the region between the olivine-wadsleyite phase boundary at a depth of 410 km, and the ringwoodite-perovskite/magnesiowüstite transition at 660 km depth, with various other phase transitions in between¹³. Transition-zone minerals have unusual properties relative to upper-mantle minerals, including water solubility¹⁴ and diffusivity of various atomic and electronic species^{15,16}. In particular, high-pressure mineral-physics studies have shown that transition-zone minerals at average mantle temperatures have anomalously high water solubility, on the order of 1 wt% (and as much as 3 wt%) whereas the solubility of water in upper- and lower-mantle minerals is less than 0.1–0.2 wt% (refs 17–19). The actual water content in the mantle can be inferred from petrological observations. In particular, volcanic rocks indicate (given estimates for their degree of melting) that the MORB source

region, which is associated with generic upper mantle, is relatively dry with 0.01 wt% water; the OIB source regions, typically presumed to reside in the lower mantle (as we also assume here), are less dry, with around 0.05 wt% water. By comparison, the upper mantle in subduction zones is most enriched in water, with approximately 0.1 wt%, probably due to dehydration of subducting slabs leading to arc volcanism²⁰.

If either or both upper and lower mantles are near chemical equilibrium with the transition zone (for example, see our model described below), and given that water solubility of the transition zone is about 10–30 times higher than that in the upper and probably lower mantles, then one can estimate a transition-zone water content between 0.1 and 1.5 wt%. Alternatively, various estimates based on planetary accretion models and geochemical mass-balance observations suggest a bulk amount of water in the entire mantle of between three and six ocean masses^{21,22}. Neither the upper nor lower mantle can be water-saturated (which would otherwise cause drastic viscosity reductions), and thus the bulk water estimates require the transition zone to have approximately 0.2–2 wt% water (such that the upper and lower mantles are between 0 and 90% of their saturation values). Thus two independent lines of reasoning lead to similar estimates for transition-zone water content that is less than the water saturation limit for transition-zone minerals (3 wt%), but probably exceeds the saturation limit for upper-mantle materials (of order 0.1 wt%; refs 17, 23).

Filtered ambient upwelling mantle

An immediate consequence of the mantle's heterogeneous water-solubility distribution is that, for a wide range of water content, vertical motion across the transition zone results in a dramatic redistribution of both water and trace elements through partial

melting. The primary vertical mantle current through which this occurs is the broad background of passively ascending ambient mantle that is forced upward by the downward injection of cold subducting slabs (it is passive in the sense that it rises more from displacement by slabs than under its own buoyancy). Given the small heat and mass flux provided by mantle plumes¹², most of the return flow from slab descent is accommodated by this background upwelling, which is itself only at ambient mantle temperatures and is upwelling at a very slow average velocity of order of magnitude $w_0 \approx 1 \text{ mm yr}^{-1}$ (see Methods). This type of circulation is common to convection models and experiments that involve a fluid layer that is heated primarily by distributed heat sources (such as radioactive isotopes), and secondarily by heating along its base¹².

In the upper transition zone, the slowly ascending ambient mantle is in the wadsleyite-dominated assembly (hereafter called 'wadsleyite' for simplicity) and, upon crossing the 410-km boundary, it transforms to an olivine-dominated assembly ('olivine' for simplicity) with significantly lower water solubility. If the wadsleyite below 410 km has water at or in excess of the solubility of olivine above 410 km, as estimated above, then once the ambient upwelling mantle material transforms to olivine it will be saturated, if not super-saturated. At 410-km-deep conditions, the ambient temperature of approximately 1,800 K (ref. 12) is greater than the eutectic point, or wet solidus, of 1,500–1,600 K (ref. 23, 24), and thus the saturated (or super-saturated) ambient mantle will undergo partial melting as soon as it crosses the 410-km boundary^{24,25}.

Given the extremely small partition coefficients (that is, high incompatibility, or preference for melt over solid) of important trace, incompatible elements, such as uranium and thorium, a melt fraction of only $f \approx 0.5$ wt% is required to reduce the concentration of these elements in the solid phase to MORB source-region

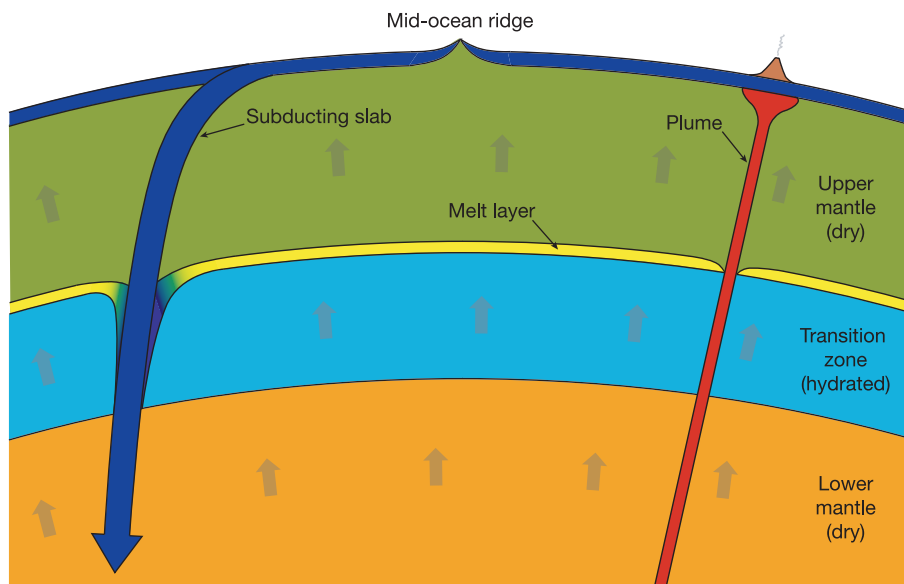


Figure 1 Sketch of the transition-zone water-filter model. Slabs subducting from cold lithosphere (dark blue) force up a broad background of passively upwelling ambient mantle (arrows) that, upon passing through the high-water-solubility transition zone (light blue) gets hydrated. When leaving the transition zone at the 410-km boundary, this ambient mantle becomes low-water solubility olivine and is thus super-saturated, wherein it partially melts, thereby extracting water and filtering off incompatible elements into the melt phase. The wet, enriched melt is likely to be heavy and thus gathers into the high-melt fraction layer trapped above the 410-km boundary (yellow). The residual solid portion of upwelling ambient mantle is buoyant but very dry and

depleted of incompatible elements; it provides the MORB source region (green). The water-filtering mechanism is suppressed in mantle plumes (red) due to the plume material's higher temperatures and velocities which result in reduced water-solubility and shorter residence times in the transition zone, thereby leading to greatly diminished hydration and thus little or no melting upon passing the 410-km boundary. Plumes thus arrive at the surface still relatively wet and enriched in compatible elements, thereby providing the source for enriched OIBs. Slabs efficiently entrain the melted material, returning water to the transition zone and incompatible elements to the deeper mantle.

concentrations, that is, a factor of 100 reduction in incompatible-element concentrations²⁶ (see Methods). Whether melting of ambient upwelling mantle reaches this limit depends on various factors, which we discuss below once the model is completely described.

After partial melting occurs at 410 km, the solid phase will continue to rise, but will be relatively dry and largely filtered of incompatible elements, thereby filling the upper mantle with the water-poor, incompatible-element-depleted material that is sampled by mid-ocean ridges. The remaining water- and incompatible-element-enriched melt residue is likely to be denser than the solid phase because of the melt-solid density crossover that is inferred to exist above the transition zone^{27,28} (see Methods). We therefore assume the melts formed at 410 km depth are denser than surrounding solid minerals in the upper mantle but lighter than transition-zone minerals^{27,29}, in which case they will accumulate and become trapped above the 410-km discontinuity.

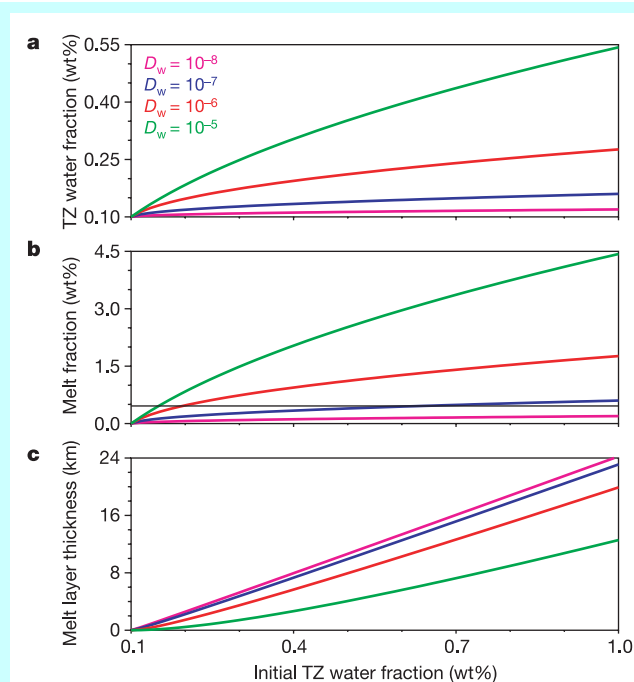


Figure 2 Results of the theoretical model. (See text and Supplementary Information.) **a**, The average water mass fraction of the transition zone (that is, in ambient mantle as it rises out of the transition zone); **b**, the melt fraction f of ambient upwelling mantle as it crosses the 410-km boundary; and **c**, the thickness of the resultant melt layer, all shown as functions of initial water content in the transition zone (that is, with no melt layer) and water/hydrogen diffusivity in slabs D_w (in $\text{m}^2 \text{s}^{-1}$). (See Supplementary section 5 for exact parameter details.) The melt fraction f needs to be above approximately 0.5 wt% (thin black line in **b**) to reduce the concentration of uranium (one of the most compatible of incompatible elements, with a partition coefficient of approximately 10^{-3}) in the remaining solid phase by factor of 100 (see Methods). Water or hydrogen diffusivity in wadsleyite at transition-zone conditions is at present unknown. Olivine hydrogen diffusivity is of the order of $10^{-8} \text{m}^2 \text{s}^{-1}$ (ref. 53); however, diffusion of larger cations in wadsleyite is faster than that in olivine by a factor of 10^3 (ref. 15). Therefore hydrogen diffusivity in wadsleyite is plausibly as much as 1,000 times faster than that in olivine; thus we consider the possible range $10^{-8} \text{m}^2 \text{s}^{-1} \leq D_w \leq 10^{-5} \text{m}^2 \text{s}^{-1}$. Higher water diffusivity yields more efficient slab entrainment, in which case only a relatively thin melt layer (that is, smaller slab-melt contact area) is required for slab entrainment to balance melt production. Thus even more water remains in the transition zone rather than the melt layer, leading to larger melt fractions in ambient mantle upwelling across the 410-km boundary. Therefore, cases with higher water diffusivity lead to thinner melt layers and larger melt fractions of ambient upwelling mantle.

Slab entrainment and water recycling

Without any sinks of mass, the trapped melt zone at 410 km would simply continue to accumulate, thicken and spread. However, once the already broad melt layer makes effective contact with cold subducting slabs, its material is readily entrained back into the deeper mantle, thereby draining the melt layer and causing a net, slow flow of melt toward slabs. In particular, a slab extending across the melt layer would cool and re-freeze silicates in its vicinity. In addition to being viscously entrained by the slab, the crystallized materials would be predominantly in the wadsleyite phase and thus also tend to sink under their own weight into the transition zone. The large water-solubility and relatively high water/hydrogen diffusivity of wadsleyite would also permit significant entrainment of water by slabs (see Supplementary section 4). Incompatible elements are also readily returned to the transition zone and lower mantle by slab entrainment; although this entails some elevation of incompatible-element concentrations in the melt layer, especially near slabs, the majority of the incompatible elements remain in the mantle below the 410-km boundary (see Supplementary section 6). Thus the slab entrainment process can, in total, return much of the enriched material back to the deeper mantle.

Most of the water carried by slabs from the melt zone is likely to be eventually deposited in the transition zone. The loss of water from slabs directly to the transition zone is dominated by hydrogen diffusion which is only moderately fast. However, after descending out of the transition zone at 660 km depth, the hydrated minerals entrained by the slab are likely to have water concentrations at the wadsleyite/ringwoodite saturation values (see Supplementary section 4) and thus well in excess of the very low water solubility of the lower mantle. Upon going through the ringwoodite-perovskite/magnesiowüstite phase transition, the slab minerals would be supersaturated and exsolve water, but not induce melting given the reduced slab temperatures and the typically high lower-mantle melting temperatures³⁰.

The resulting porosity of the matrix holding the exsolved water would be approximately $\rho_s X_{wd}^* / \rho_w = 6\%$ (where $\rho_s \approx 4,000 \text{kg m}^{-3}$ is silicate density, $\rho_w \approx 2,000 \text{kg m}^{-3}$ is approximate water density at 660 km depth³¹, and $X_{wd}^* = 3 \text{wt}\%$ is the wadsleyite/ringwoodite water saturation mass fraction at slab temperatures) and thus probably of high enough permeability to allow water to percolate back up into the transition zone, where it would be re-absorbed. The return of water to the transition zone by either diffusion or percolation would also be facilitated by the tendency of many slabs to flatten out, permanently or temporarily, in the transition zone at the 660-km boundary^{1,2}. Therefore, the production of heavy melts above 410 km and water exsolution beneath 660 km would conceivably lead to water being trapped in the transition zone; without these effects, water could instead be convectively well-mixed over the entire mantle³².

A simple circulation model

With the inferred trapping and closed circulation of water through the transition zone and overlying melt layer, we can prescribe a simple theoretical model with which to make basic predictions about water concentrations in the transition zone, the melt fraction of ambient mantle upwelling across the 410-km boundary, and the average thickness of the melt layer (Fig. 2). The model is driven by a prescribed background flow of downwelling slabs and the resultant passive ambient upwelling. Given the temperature and water concentration of ambient upwelling, the fraction of melt produced when material crosses the 410-km boundary can be calculated. The flux of silicates, water and incompatible elements injected into the heavy melt layer can then be balanced against their eventual entrainment by slabs back into the transition zone and lower mantle (assuming that all entrained water is deposited in the transition

zone), where they are then recirculated back into the ambient upwelling mantle. See Supplementary Information for model details.

Results of the model (Fig. 2) are cast as functions of the two most poorly constrained parameters: the initial water concentration in the transition zone (that is, before formation of a melt layer), and water diffusivity in wadsleyite, which controls slab entrainment of melt-layer material (see Fig. 2 legend and Supplementary section 5). The model predicts that water in the transition zone is always reduced below its initial value (because it must lose water to the melt layer) and thus remains well below saturation values. The melt fraction of ambient upwelling mantle exceeds the minimum necessary for filtering off incompatible elements (0.5 wt%) if wadsleyite water diffusivity is more than ten times that for olivine, which is plausible. The melt layer thickness varies between 1 km and of the order of 10 km; those cases which have the largest melt fraction and thus best filter off incompatible elements (cases with high water diffusivity) are also associated with the narrowest melt layers (see Fig. 2 legend).

Unfiltered plumes and the Archaean mantle

In contrast to upwelling ambient mantle, the water-filter effect would be largely suppressed in mantle plumes, thus allowing them to deliver moderately enriched mantle material of OIB chemistry to the surface. In particular, upon passing through the transition zone, plume material would conceivably absorb less water than would ambient upwelling mantle.

First, plume material moves with an average velocity on the order

of at least $w_{\text{plume}} = 100 \text{ cm yr}^{-1}$ (ref. 12). The zone of water entrainment by hydrogen diffusion into the plume is approximated by $\sqrt{D_w H / w_{\text{plume}}}$ where D_w is hydrogen diffusivity in the transition zone, and $H = 250 \text{ km}$ is the transition zone thickness. Given a possible range of diffusivity of $10^{-8} \leq D_w \leq 10^{-5} \text{ m}^2 \text{ s}^{-1}$ (see Fig. 2 legend), the entrainment zone is likely to be between 300 m and 10 km, which implies that most of the plume (with typical radii of 100 km) remains relatively dry.

Second, recent experimental studies suggest that while water solubility increases with temperature in olivine³³, it decreases with temperature in transition-zone minerals wadsleyite¹⁴ and ringwoodite³⁴. Consequently, hot plume material (with temperatures around 2,100 K) passing through the transition zone would have little capacity for water, and would probably be well under-saturated when transforming to olivine above 410 km (see Methods), thus precluding any dehydration melting. Mantle plumes would therefore neither be filtered of their incompatible elements nor stripped of the little water they have. Eventually, upon approaching the surface, the 'damp' plumes would undergo pressure-release melting, leading to hotspots that are wet relative to the much drier MORB, and with enriched OIBs. For similar reasons, the water-filter mechanism could have been suppressed during the Archaean period when temperatures in the Earth were perhaps 200 K higher than today³⁵ resulting in low transition-zone water concentrations, and thus little or no dehydration-induced melting of mantle material upwelling across the 410 km boundary (Methods). Most of the highly enriched continental crust could therefore have been extracted from the Archaean mantle³⁶ without impedance from the water-filter mechanism.

Geochemical consequences

In addition to their different concentrations of incompatible elements, OIB and MORB source regions also have significantly different isotope ratios (such as between concentrations of radioactive daughter products such as ^{206}Pb to original or primordial ones like ^{204}Pb). The distribution of isotope ratios for OIB is usually much broader than that for MORB, although the two distributions often overlap or are proximal to one another in isotope-ratio space²⁶. The difference in distributions is usually attributed to OIB and MORB source regions evolving isotopically in isolation from each other for a billion years or so after formation of the core and crust²⁶. However, in our model (Fig. 3), this difference can be attributed to mixing and homogenization of isotope ratios in ambient mantle passing through the hypothetical 410 km melt zone (which causes the MORB distribution to be narrower than, and initially overlap with, the OIB distribution), along with possible drift between the MORB and OIB distributions over a few hundred million to a billion years (also depending on the relative fractionation of parent and daughter isotopes; see Fig. 3).

The abundances of noble gases, such as He and Ar, and their isotopes are also affected by the hypothetical partial melting near 410 km. These elements probably behave like all other highly incompatible elements^{37,38} and hence would tend to be sequestered in the deep mantle by the proposed filter mechanism. Therefore the inferred enrichment of the radiogenic isotope ^{40}Ar , as well as its parent ^{40}K , in the deep mantle relative to the MORB source region³⁹, is naturally explained by our model. The amount of atmospheric ^{40}Ar , which is larger than what can have been generated by the time-integrated amount of ^{40}K in the crust and MORB source region, may be provided by flux of argon from the deep mantle via plumes, as well as extraction from the greater part of the mantle (including deeper ^{40}Ar -enriched portions) during the Archaean⁴⁰, before the filter mechanism was activated (see previous section).

The low flux of ^4He relative to the global heatflow⁴¹ may be similarly explained in terms of helium incompatibility and partial sequestration in the lower mantle by the filter mechanism. However,

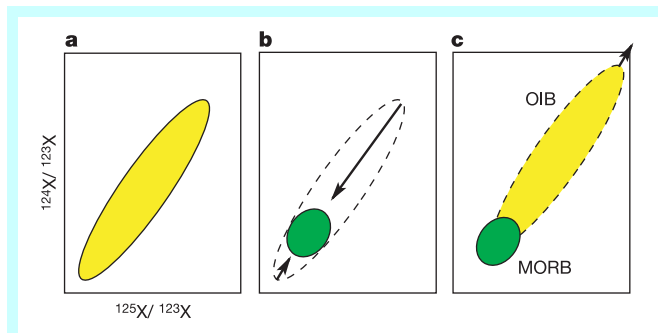


Figure 3 Schematic diagram of MORB and OIB isotopic variation according to the water-filter model. ^{124}X and ^{125}X represent concentrations of generic radiogenic isotopes (such as ^{206}Pb and ^{207}Pb), whereas ^{123}X applies to non-radiogenic ones (such as ^{204}Pb). **a**, Isotopic compositions in mantle below 410 km are heterogeneous, typical for the OIB source region. **b**, Partial melting of ambient mantle above 410 km leads to mixing of these heterogeneous materials in the melt layer; chemical equilibration of the well-mixed melt with the solid phase homogenizes the isotopic composition of the solid to a narrow distribution. The resulting isotopic composition of the solid is a weighted average of isotopic compositions in the deeper region; assuming that either FOZO^{4,54} or the Hawaii source region are volumetrically dominant in the deep mantle, the weighted average will be near the depleted end-member composition (**b**). Above 410 km, the depleted solid materials can undergo isotopic evolution for several hundred million to a billion years (given the average 1 mm yr^{-1} order-of-magnitude velocity of ambient mantle) before reaching mid-ocean ridges. **c**, During this period, the isotopic compositions of the MORB and OIB source materials will shift, leading to present-day observations. The sense of shift is determined by the incompatibilities of parent and daughter elements. If the parent is more incompatible than the daughter (as with U–Pb and Rb–Sr systems)⁵⁵ then the MORB source will be relatively depleted in parent elements and undergo less isotopic evolution than will the OIB source (**c**, shift indicated by arrow). However, when the daughter is more incompatible than the parent (as in the Sm–Nd system⁵⁵, and possibly the U–He and Th–He systems^{37,38}) the opposite trend occurs; the MORB source will grow more enriched in radiogenic isotopes relative to nonradiogenic ones, which may explain the high $^4\text{He}/^3\text{He}$ for MORB relative to that for OIB).

one major limitation of these arguments is the large uncertainty in the partition coefficients of many incompatible elements, especially noble gases, between melts and solids³⁸. Moreover, the validity of assumptions behind certain noble-gas constraints can be challenged; for example, assumptions about the K/U ratio in the bulk silicate Earth (which is used to estimate the amount of missing ⁴⁰Ar and ⁴⁰K), and the steady-state assumption used in the various noble-gas flux arguments have been recently scrutinized^{42,43}.

Conclusion

The model presented herein suggests a new form of whole-mantle flow wherein the bulk circulation of mantle material is, via the transition-zone water filter, decoupled from the circulation of incompatible elements. The model therefore explains much of the observations of geochemical reservoirs, including OIB and MORB source regions as well as continental crust, but without invoking layered convection and all its attendant problems. Although our model predicts that the deep mantle is enriched with radioactive elements, the style of convection will be markedly different from that of a layered model which precludes mass transfer between layers. In our model, major-element mass transfer occurs between the two layers, and so the heat generated by radioactive elements in the deep mantle is carried away by a whole-mantle-scale circulation.

One of the predictions of this model is the presence of a layer of melt or partial melt above the 410-km discontinuity, which has also been suggested by other workers²⁴ but only for the pre-Mesozoic Earth. Obviously, global seismic models^{8,9} have not resolved such a melt zone and thus the layer would have to be relatively thin, say less than of the order of 10 km, which is in keeping with the predictions of our model calculations (see Fig. 2). The existence of a melt layer above the 410-km discontinuity has been suggested by recent seismological studies^{44,45}, although these were focused on particular regions and thus the presence of melt over greater areas is unknown. Further high-frequency body-wave studies and additional coverage offered by large-area land-based and ocean-bottom passive arrays could elucidate whether a thin melt layer above 410 km exists broadly.

Our hypothesis also requires further testing. In particular, aspects of our model hinge on several mineral physics issues which remain highly uncertain. These include the density of hydrous melts at high pressures; diffusivity of hydrogen in transition-zone minerals; water solubility in lower-mantle minerals; temperature-dependence of water solubility in transition-zone minerals; and the partition coefficients of incompatible elements between melt and silicate minerals at transition-zone and deep upper-mantle conditions.

Moreover, sophisticated convection models incorporating the proposed additional physics should shed further light on the feasibility of the mechanism, but will also require considerable technical innovation to treat not only the motion and heat transport of bulk mantle, but also transport of incompatible (most importantly heat-producing) elements, water and partial melts. For now, we leave the transition-zone water filter as an alternative hypothesis to layered convection that can be elaborated on and tested by future geodynamical, seismological, mineralogical and geochemical studies. □

Methods

Rise velocity of ambient upwelling mantle

An important controlling parameter in the water-filter model is the ascent velocity of upwelling ambient mantle as it passes out of the transition zone. This velocity determines, for example, both the rate of production of melts at the 410-km boundary, as well as the transit time for material crossing the upper mantle. Mass conservation across the 410-km boundary requires that upwelling and downwelling volume fluxes balance, that is $2\pi R d_{\text{slab}} w_{\text{slab}} \approx 4\pi R^2 w_0$ where $R \approx 6,000$ km is the radius of the 410-km boundary, $d_{\text{slab}} \approx 100$ km is a typical slab thickness, $w_{\text{slab}} = 10 \text{ cm yr}^{-1}$ is the order-of-magnitude for a typical slab descent velocity, and w_0 is the average ascent velocity of ambient upwelling mantle. We have assumed that the net plume flux makes a minor contribution and that the net length of the intersection of all slabs with the 410-km boundary is approximately $2\pi R$ because most slabs occur in a nearly great circle around the Pacific rim (and any additional slabs largely compensate for the Pacific rim being slightly smaller than

a great circle). This leads to an order-of-magnitude estimate of $w_0 \approx 1 \text{ mm yr}^{-1}$, although actual velocities are probably between this and as much as a factor of two smaller.

Melting at the 410-km boundary

Two essential ingredients are needed for the water-filter mechanism to work: (1) a degree of melting of ambient upwelling mantle (as it crosses the 410-km boundary) that is sufficient to filter off incompatible elements; and (2) a resulting melt phase that is heavier than the solid (in order for the melt residue to become trapped at the 410-km boundary until it is entrained by slabs back into the deeper mantle).

Although the melt is presumed heavy and the solid ascends slowly, the diffusivities of incompatible elements such as uranium and thorium are extremely small; thus the melting can be treated as fractional (that is, the phases separate faster than their concentrations of incompatible elements can reach chemical equilibrium). In this case the ratio of incompatible element concentration in the solid phase after partial melting (C_s) to that prior to melting (C_{s0}) is given by $C_s/C_{s0} = (1 - f)^{D-1}$ (ref. 46) where D is the partition coefficient between solid and melt. Major incompatible elements, such as uranium and thorium, have $D \approx 10^{-3}$, so to reduce major incompatible elements by a factor of 100 (as is required for the depletion of the MORB source region²⁶) one must have a melt fraction of $f \approx 0.5 \text{ wt\%}$.

The assumption that water- and trace-element-enriched melt is denser than the solid phase is based on the melt-solid density crossover that is inferred to exist in the upper mantle^{27–29}. The density crossover, however, has only been investigated for anhydrous melts at high temperatures (for example, $\sim 2,300$ K as is relevant for 410 km depths)²⁹. The effect of water addition (approximately 10 wt% in the melt²³; see also Supplementary section 2) on the crossover is at present unknown, but is likely to have a few competing influences. The purely chemical effect of water addition (at the concentrations inferred here) is to reduce melt density; however, since water dissolves in the melt as OH^- as well as molecular water^{47–49} the density reduction is relatively small^{49,50}; for our model, at most $\sim 5\%$ (calculated from experimental data^{49,50}). An equally important effect of water addition is the reduction of melting temperature^{23,24}. In particular, the 1,800 K hydrous melt in our model is 500 K cooler than dry melt at equivalent pressure; given that melt thermal expansivity is much larger than that of solids^{51,52}, the temperature reduction is likely to cause at least a $\sim 5\%$ increase in the melt-solid density difference (relative to 2,300 K conditions). The combined effects are likely to offset each other such that the density crossover occurs in the upper mantle even for our hydrous melt.

Water solubility above and below 410 km

The operation of the water-filter mechanism also depends on the contrast in water-solubility across the 410 km boundary. If the water-solubility of warm or hot upwelling mantle (hotter than the wet-olivine solidus temperature of 1,500–1,600 K) is significantly greater below the 410-km boundary than above it, then the upwelling material can have a wide range of water concentrations that guarantee it will undergo partial melting and subsequent filtering upon crossing the boundary. If the solubility is instead less beneath 410-km than above, then melting is not guaranteed and may even be precluded. The solubility limit X^* for a given mineral depends on temperature T and pressure P as:

$$X^*(T, P) = A\varphi(T, P)e^{-\frac{E_a + PV_a}{RT}} \quad (1)$$

where A is a constant, φ is water fugacity, E_a and V_a are the activation energy and volume, respectively, for dissolution of water. All the parameters have been determined for olivine^{17,33}, and the temperature dependence of water solubility at 410 km can be estimated. The results show that the water solubility in olivine increases with temperature, that is, from less than $<0.1 \text{ wt\%}$ at 1,400 K to $\sim 1 \text{ wt\%}$ at 2,200 K; at ambient mantle temperatures of 1,800 K this solubility is $\sim 0.2 \text{ wt\%}$. The temperature dependence of water solubility in transition-zone minerals is not well known. However, experimental results on ringwoodite³⁴ show water solubility decreasing with temperature, from $\sim 3 \text{ wt\%}$ at 1,400 K to $\sim 0.2 \text{ wt\%}$ at 2,200 K; it is approximately $\sim 1 \text{ wt\%}$ at 1,800 K. Results for wadsleyite are less robust but a recent compilation of existing data¹⁴ suggests a similar trend: water solubility decreases with temperature. Assuming that the temperature dependence of water solubility in wadsleyite is similar to that of ringwoodite (as with many other properties^{15,16} including water-solubility itself), the temperature at which the water solubility in wadsleyite becomes the same as that in olivine is approximately 2,000 K, typical for plumes and Archaean mantle. Thus while there are sufficient conditions for melting of warm present-day ambient mantle, the same is not true of hotter plumes and Archaean mantle. However, melting of mantle upwellings is only precluded if their water concentrations are less than the solidus water concentration, which is generally smaller than the olivine solubility limit. Assuming plume material arises from a relatively dry lower mantle, its water concentrations can remain low enough to avoid melting at 410 km because of the combination of its low water-solubility and short residence time in the 'wet' transition zone. Low water concentrations in the hotter Archaean mantle would conceivably occur because the reduced solubility of the transition zone precludes it from being a water trap, thus distributing water across the mantle.

Received 11 July 2002; accepted 18 July 2003; doi:10.1038/nature01918.

- Grand, S. Mantle shear structure beneath the Americas and surrounding oceans. *J. Geophys. Res.* **99**, 11591–11621 (1994).
- van der Hilst, R., Widiyantoro, S. & Engdahl, E. Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578–584 (1997).
- Masters, G., Laske, G., Bolton, H. & Dziewonski, A. in *Earth's Deep Interior* (eds Karato, S., Forte, A., Liebermann, R., Masters, G. & Stixrude, L.) 63–87 (AGU, Washington DC, 2000).
- van Keken, P., Hauri, E. & Ballentine, C. Mantle mixing: The generation, preservation, and destruction of chemical heterogeneity. *Annu. Rev. Earth Planet. Sci.* **30**, 493–525 (2002).
- Hofmann, A. & White, W. The role of subducted oceanic crust in mantle evolution. *Carnegie Inst. Wash. Yearbk.* **79**, 477–483 (1980).

6. Coltice, N. & Ricard, Y. Geochemical observations and one layer mantle convection. *Earth Planet. Sci. Lett.* **174**, 125–137 (1999).
7. Kellogg, L., Hager, B. & van der Hilst, R. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
8. Dziewonski, A. & Anderson, D. Preliminary reference earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
9. Kennet, B. & Engdahl, E. Travel times for global earthquake location and phase identification. *Geophys. J. Int.* **105**, 429–466 (1991).
10. Davaille, A. Simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature* **402**, 756–760 (1999).
11. Tackley, P. Strong heterogeneity caused by deep mantle layering. *Geochem. Geophys. Geosyst.* **3**, 10.1029/2001GC000167 (2002).
12. Schubert, G., Turcotte, D. & Olson, P. *Mantle Convection in the Earth and Planets* (Cambridge Univ. Press, Cambridge, UK, 2001).
13. Ringwood, A. Phase transformations and their bearings on the constitution and dynamics of the mantle. *Geochem. Cosmochem. Acta* **55**, 2083–2110 (1991).
14. Williams, Q. & Hemley, R. Hydrogen in the deep earth. *Annu. Rev. Earth. Planet. Sci.* **29**, 365–418 (2001).
15. Chakraborty, S. *et al.* Enhancement of cation diffusion rates across the 410-kilometer discontinuity in earth's mantle. *Science* **283**, 362–365 (1999).
16. Xu, Y., Poe, B., Shankland, T. & Rubie, D. Electrical conductivity of olivine, wadsleyite, and ringwoodite under upper-mantle conditions. *Science* **280**, 1415–1418 (1998).
17. Kohlstedt, D., Keppler, H. & Rubie, D. Solubility of water in the α , β and γ phases of (Mg, Fe)₂SiO₄. *Contrib. Mineral. Petrol.* **123**, 345–357 (1996).
18. Bolfan-Casanova, N., Keppler, H. & Rubie, D. Water partitioning between nominally anhydrous minerals in the MgO–SiO₂–H₂O system up to 24 GPa: implications for the distribution of water in the earth's mantle. *Earth Planet. Sci. Lett.* **182**, 209–221 (2000).
19. Murakami, M., Hirose, K., Yurimoto, H., Nakashima, S. & Takafuji, N. Water in earth's lower mantle. *Science* **295**, 1885–1887 (2002).
20. Sobolev, A. & Chaussidon, M. H₂O concentration in primary melts from supra-subduction zones and mid-ocean ridges: implications for H₂O storage and recycling in the mantle. *Earth Planet. Sci. Lett.* **137**, 45–55 (1996).
21. Ringwood, A. *Composition and Structure of the Earth's Mantle* (McGraw-Hill, New York, 1975).
22. Ahrens, T. Water storage in the mantle. *Nature* **342**, 122–123 (1989).
23. Inoue, T. Effect of water on melting phase relations and melt composition in the system Mg₂SiO₄–MgSiO₃–H₂O up to 15 GPa. *Phys. Earth. Planet. Inter.* **85**, 237–263 (1994).
24. Kawamoto, T., Hervig, R. & Holloway, J. Experimental evidence for a hydrous transition zone in the early earth's mantle. *Earth Planet. Sci. Lett.* **142**, 587–592 (1996).
25. Young, T., Green, H., Hofmeister, A. & Walker, D. Infrared spectroscopic investigation of hydroxyl in β –(Mg, Fe)₂SiO₄ and coexisting olivine: implications for mantle evolution and dynamics. *Phys. Chem. Minerals* **19**, 409–422 (1993).
26. Hofmann, A. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–228 (1997).
27. Stolper, E., Walker, D., Hager, B. & Hays, J. Melt segregation from partially molten source regions: the importance of melt density and source region size. *J. Geophys. Res.* **86**, 6261–6271 (1981).
28. Ohtani, E. & Maeda, M. Density of basaltic melt at high pressure and stability of the melt at the base of the lower mantle. *Earth Planet. Sci. Lett.* **193**, 69–75 (2001).
29. Ohtani, E., Nagata, Y., Suzuki, A. & Kato, T. Melting relations of peridotite and the density crossover in planetary mantles. *Chem. Geol.* **120**, 207–221 (1995).
30. Boehler, R. Melting temperature of the earth's mantle and core: Earth's thermal structure. *Annu. Rev. Earth Planet. Sci.* **24**, 15–40 (1996).
31. Pitzer, K. & Sterner, S. Equations of state valid continuously from zero to extreme pressures for H₂O and CO₂. *J. Chem. Phys.* **101**, 3111–3116 (1994).
32. Richard, G., Monnerau, M. & Ingrin, J. Is the transition zone an empty water reservoir? Inferences from numerical model of mantle dynamics. *Earth Planet. Sci. Lett.* **205**, 37–51 (2002).
33. Zhao, Y.-H., Ginsberg, S. & Kohlstedt, D. Experimental investigation on water solubility in olivine single crystals with different Fe content. *Acta Petrol. Sinica* **17**, 123–128 (2001).
34. Ohtani, E., Mizobata, H. & Yurimoto, H. Stability of dense hydrous magnesium silicate phases in the system Mg₂SiO₄–H₂O and MgSiO₃–H₂O at pressures up to 27 GPa. *Phys. Chem. Minerals* **27**, 533–544 (2000).
35. Richter, F. Models of Archean thermal regimes. *Earth Planet. Sci. Lett.* **73**, 350–360 (1985).
36. Taylor, S. & McLennan, S. *The Continental Crust: Its Composition and Evolution* (Blackwell, Oxford, UK, 1985).
37. Carroll, M. & Stolper, E. Noble gas solubilities in silicate melts and glasses: new experimental results for argon and relationship between solubility and ionic porosity. *Geochem. Cosmochem. Acta* **57**, 5039–5051 (1993).
38. Chamorro, E. *et al.* Ar and K partitioning between clinopyroxene and silicate melt to 8 GPa. *Geochem. Cosmochem. Acta* **66**, 507–519 (2002).
39. Allège, C., Hofmann, A. & O'Nions, K. The argon constraints on mantle structure. *Geophys. Res. Lett.* **23**, 3555–3557 (1996).
40. Hamano, Y. & Ozima, M. in *Terrestrial Rare Gases* (eds Alexander, E. & Ozima, M.) 155–171 (Japan Scientific Society Press, Tokyo, 1978).
41. O'Nions, R. & Oxburgh, E. Heat and helium in the earth. *Nature* **306**, 429–431 (1983).
42. Albarède, F. Time-dependent models of U–Th–He and K–Ar evolution and the layering of mantle convection. *Chem. Geol.* **145**, 413–429 (1998).
43. Ballentine, C., Van Keken, P., Porcelli, D. & Hauri, E. Numerical models, geochemistry, and the Zero Paradox noble gas mantle. *Phil. Trans. R. Soc. Lond.* **360**, 2611–2631 (2002).
44. Revenaugh, J. & Sipkin, S. Seismic evidence for silicate melt atop the 410-km mantle discontinuity. *Nature* **369**, 474–476 (1994).
45. Vinnik, L. & Farra, V. Subcratonic low-velocity layer and flood basalts. *Geophys. Res. Lett.* **29** doi: 10.1029/2001GL014064 (2002).
46. Shaw, D. Trace element fractionation during anatexis. *Geochem. Cosmochem. Acta* **34**, 237–243 (1970).
47. Burnham, C. in *The Evolution of the Igneous Rocks* (ed. Yoder, J.) 439–482 (Princeton Univ. Press, Princeton, NJ, 1979).
48. Silver, L. & Stolper, E. A thermodynamic model for hydrous silicate melts. *J. Geol.* **93**, 161–178 (1985).
49. Ochs, F. & Lange, R. The partial molar volume, thermal expansivity, and compressibility of H₂O in NaAlSi₃O₈ liquid. *Contrib. Mineral. Petrol.* **129**, 155–165 (1997).
50. Richet, P. *et al.* Water and the density of silicate glasses. *Contrib. Mineral. Petrol.* **138**, 337–347 (2000).
51. Fei, Y. in *Mineral Physics and Crystallography A Handbook of Physical Constants* (ed. Ahrens, T.) AGU Ref. Shelf 2, 29–44 (AGU, Washington DC, 1995).
52. Suzuki, A., Ohtani, E. & Kato, T. Density and thermal expansion of a peridotite melt at high pressure. *Phys. Earth Planet. Inter.* **107**, 53–61 (1998).
53. Mackwell, S. & Kohlstedt, D. Diffusion of hydrogen in olivine: implications for water in the mantle. *J. Geophys. Res.* **95**, 5079–5088 (1990).
54. Hart, S., Hauri, E., Oschmann, L. & Whitehead, J. Mantle plumes and entrainment: isotopic evidence. *Science* **256**, 517–520 (1992).
55. Hauri, E., Wagner, T. & Grove, T. Experimental and natural partitioning of Th, U, Pb and other trace elements between garnet, clinopyroxene and basaltic melt. *Chem. Geol.* **117**, 149–166 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Suzuki and E. Ohtani for access to data and unpublished documents, and G. Leahy for help producing Fig. 1. We also thank A. Hofmann for reviewing the manuscript and R. Batiza, T. Grove, S. Hart, E. Hauri, M. Hirschman, A. Rempel, Y. Ricard, F. Richter, G. Schubert, N. Sleep, D. Stevenson, K. Turekian, L. Vinnik, W. White, and D. Yuen for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.B. (david.bercovici@yale.edu).

Violation of a Bell-like inequality in single-neutron interferometry

Yuji Hasegawa¹, Rudolf Loidl^{1,2}, Gerald Badurek¹, Matthias Baron^{1,2} & Helmut Rauch¹

¹Atominstitut der Österreichischen Universitäten, Stadionallee 2, A-1020 Wien, Austria

²Institute Laue Langevin, B. P. 156, F-38042 Grenoble Cedex 9, France

Non-local correlations between spatially separated systems have been extensively discussed in the context of the Einstein, Podolsky and Rosen (EPR) paradox¹ and Bell's inequalities². Many proposals and experiments designed to test hidden variable theories and the violation of Bell's inequalities have been reported^{3–7}; usually, these involve correlated photons, although recently an experiment was performed with ⁹Be⁺ ions⁸. Nevertheless, it is of considerable interest to show that such correlations (arising from quantum mechanical entanglement) are not simply a peculiarity of photons. Here we measure correlations between two degrees of freedom (comprising spatial and spin components) of single neutrons; this removes the need for a source of entangled neutron pairs, which would present a considerable technical challenge. A Bell-like inequality is introduced to clarify the correlations that can arise between observables of otherwise independent degrees of freedom. We demonstrate the violation of this Bell-like inequality: our measured value is 2.051 ± 0.019 , clearly above the value of 2 predicted by classical hidden variable theories^{9–12}.

The concept of quantum non-contextuality^{9–12} represents a straightforward extension of the classical view: the result of a particular measurement is determined independently of previous (or simultaneous) measurements on any set of mutually commuting observables¹³. Local theories represent a particular circumstance of non-contextuality, in that the result is assumed not to depend on measurements made simultaneously on spatially separated (mutually non-interacting) systems. In order to test non-contextuality, joint measurements of commuting observables that are not necessarily separated in space are required.

Two degrees of freedom can be used for such experiments, for example, the spatial and the spinor properties, of single particles, prepared in an a non-factorized state and manipulated to measure two commuting observables. A Bell-like inequality has been obtained to distinguish non-contextual hidden variable (NCHV) theories from the prediction of quantum mechanics¹⁴.

Here, we report a single-neutron interferometer experiment to show stronger correlations than the classical non-contextual model with the use of a Bell-like inequality. (General descriptions of neutron interferometer experiments are summarized in the literature¹⁵.) We note that all the experiments showing the violation of the Bell's inequalities have been performed with correlated entangled pairs^{16–20}, including the recent one with ⁹Be⁺ ions⁸. We used not entangled pairs but single neutrons and the entanglement is achieved between different degrees of freedom in a single particle. This is based on the fact that states of spin-1/2 particles, like neutrons, are described by a tensor product Hilbert space, that is, $H = H_1 \otimes H_2$ where H_1 and H_2 are respectively disconnected Hilbert spaces corresponding to the spatial and the spinor wave function. Observables of the spatial part are commutable with those of the spinor part, this justifying the derivation of a Bell-like inequality by the NCHV theories¹⁴. The experiment consists of joint measurements of commuting observables of single neutrons in an appropriately prepared nonfactorizable state.

In our polarized neutron interferometer experiment, the total wavefunction consists of the entanglement of the spatial part and

the spinor part²¹, that is, different degrees of freedom. The normalized total wavefunction $|\Psi\rangle$ can be represented as a Bell state, $|\Psi\rangle = \frac{1}{\sqrt{2}}(|\downarrow\rangle \otimes |I\rangle + |\uparrow\rangle \otimes |II\rangle)$. Here, $|\uparrow\rangle$ and $|\downarrow\rangle$ denote the up-spin and down-spin states, and $|I\rangle$ and $|II\rangle$ denote the two beam paths in the interferometer.

The expectation value for the joint measurement for the spinor $\frac{1}{\sqrt{2}}(|\uparrow\rangle + e^{i\alpha}|\downarrow\rangle)$ and the path $\frac{1}{\sqrt{2}}(|I\rangle + e^{i\chi}|II\rangle)$ is calculated to be:

$$E'(\alpha, \chi) = \langle \Psi | \hat{P}^s(\alpha) \cdot \hat{P}^p(\chi) | \Psi \rangle \\ = \langle \Psi | [(+1) \cdot \hat{P}_{\alpha+1}^s + (-1) \cdot \hat{P}_{\alpha-1}^s] \\ [(+1) \cdot \hat{P}_{\chi+1}^p + (-1) \cdot \hat{P}_{\chi-1}^p] | \Psi \rangle \quad (1)$$

where $\hat{P}_{\alpha\pm 1}^s$ and $\hat{P}_{\chi\pm 1}^p$ are the projection operators to the states $\frac{1}{\sqrt{2}}(|\uparrow\rangle \pm e^{i\alpha}|\downarrow\rangle)$ and $\frac{1}{\sqrt{2}}(|I\rangle \pm e^{i\chi}|II\rangle)$, respectively. (These projection operators and the expectation value correspond to $P_{\pm}(\mathbf{a})$, $P_{\pm}(\mathbf{b})$ and $E(\mathbf{a}, \mathbf{b})$ in the conventional EPR argument²².) It should be emphasized here that the observables $\hat{P}^s$ and $\hat{P}^p$ operate in different Hilbert spaces, and thus commute each other.

A Bell-like inequality for a single-neutron experiment is expressed¹⁴ with the expectation values $E'(\alpha, \chi)$ as $-2 \leq S' \leq 2$, with $S' \equiv E'(\alpha_1, \chi_1) + E'(\alpha_1, \chi_2) - E'(\alpha_2, \chi_1) + E'(\alpha_2, \chi_2)$.

In our experiments, the expectation value $E'(\alpha, \chi)$ has been determined by a combination of count rates in a single detector with appropriate setting of χ and α . This is given by:

$$E'(\alpha, \chi) \\ = \frac{N'(\alpha, \chi) + N'(\alpha + \pi, \chi + \pi) - N'(\alpha, \chi + \pi) - N'(\alpha + \pi, \chi)}{N'(\alpha, \chi) + N'(\alpha + \pi, \chi + \pi) + N'(\alpha, \chi + \pi) + N'(\alpha + \pi, \chi)} \quad (2)$$

where $N'(\alpha_j, \chi_k)$ denotes the count rate with the spin-rotation of α_j and the phase shift of χ_k , which is given by $N'(\alpha_j, \chi_k) = \langle \Psi | \hat{P}_{\alpha_j+1}^s \cdot \hat{P}_{\chi_k+1}^p | \Psi \rangle$. This accounts for measurements to determine the expectation value of the joint measurement with successive counts in one detector.

Quantum theory predicts sinusoidal behaviour for the count rate $N'^{\text{qm}}(\alpha, \chi) = \frac{1}{2}\{1 + \cos(\alpha + \chi)\}$. The same behaviour is also expected for the expectation value $E'(\alpha, \chi) = \cos(\alpha + \chi)$. These functions will show the violation of the Bell-like inequality for various sets of the polarization analysis (α) and the phase shift (χ). The maximum violation is expected, for instance, for the set, $\alpha_1 = 0$, $\alpha_2 = \pi/2$, $\chi_1 = \pi/4$, and $\chi_2 = -\pi/4$, as $S' = 2\sqrt{2} = 2.82 > 2$. So far we have described the experiment in terms of perfect implementation. In the actual experiment, however, perfect sinusoidal dependence of N' cannot be established owing to unavoidable component misalignments, imperfect quality of polarization/interference, and so on, which is characterized by contrasts of the oscillations. The value, S' , reduces in proportion to these contrasts. Thus, average contrasts of more than 70.7% ($= \sqrt{2}/2$) is essential in order to show the violation of the Bell-like inequality.

Our experiments consist of three stages: preparation, manipulation and detection. The preparation was achieved by the use of a spin-turner after the polarized beam was split into two, producing a Bell state, $|\Psi_{\text{exp}}\rangle = \frac{1}{\sqrt{2}}(|\rightarrow\rangle \otimes |I\rangle + |\leftarrow\rangle \otimes |II\rangle)$. In the manipulation, two parameters of χ and α were adjusted. And finally, neutrons with certain properties were detected. In recent work, four detectors were used to measure the total outcomes of Stern–Gerlach polarization analysers installed in both interfering beams¹⁴. But it is shown that polarization analysis of the O-beam is sufficient to obtain correlation coefficients.

The experiment was carried out at the silicon-perfect-crystal interferometer beam line S18 at the high flux reactor at the Institute Laue Langevin²³. A schematic view of the experimental set-up is shown in Fig. 1. The neutron beam was monochromatized to have a

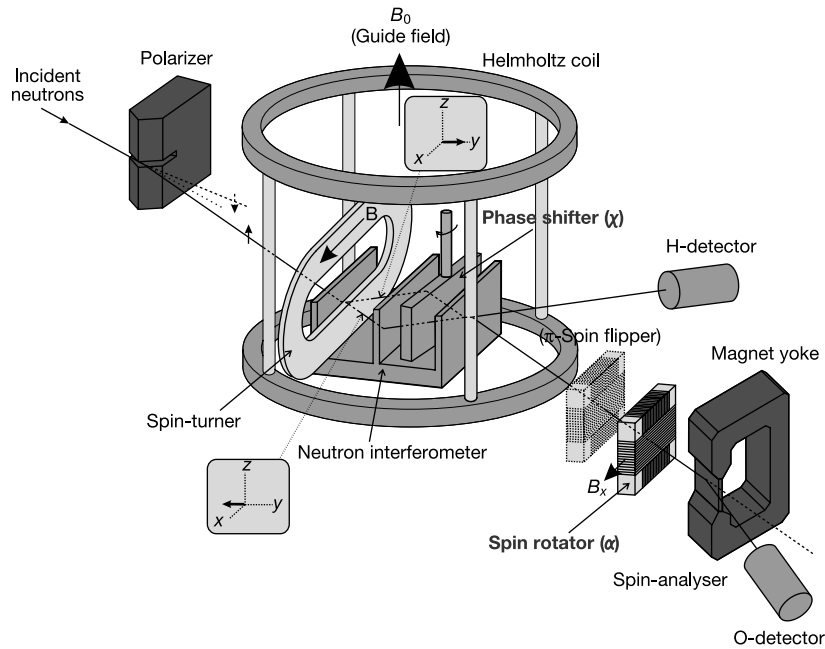


Figure 1 Schematic view of the experimental set-up to observe quantum correlations clarified by the Bell-like inequality in single-neutron interferometry. The experiments consists of three processes. (1) Preparation of the entangled state $|\Psi_{\text{exp}}\rangle = \frac{1}{\sqrt{2}}(|\rightarrow\rangle\otimes|I\rangle + |\leftarrow\rangle\otimes|II\rangle)$: an incident neutron beam is polarized with the use of a magnetic-prism polarizer. This up-polarized beam falls on the silicon-perfect-crystal neutron interferometer and splits into two beam paths, $|I\rangle$ and $|II\rangle$. A Mu-metal

spin-turner directs $|\uparrow\rangle$ to $|\rightarrow\rangle$ for one beam and to $|\leftarrow\rangle$ for the other. (2) Manipulation of the two parameters: a phase shift χ and a spinor rotation angle α are adjusted. Together with a Heusler spin-analyser, neutrons with particular spinor and path properties are selected, resulting in $\hat{\rho}_{\alpha+1}^s \cdot \hat{\rho}_{\chi+1}^p |\Psi\rangle$. (3) Detection: numbers of neutrons are counted, yielding $N'(\alpha, \chi) = \langle \Psi | \hat{\rho}_{\alpha+1}^s \cdot \hat{\rho}_{\chi+1}^p | \Psi \rangle$.

mean wave length of $\lambda_0 = 1.92(2) \text{ \AA}$ by the use of a silicon perfect crystal monochromator. This incident beam was polarized vertically by magnetic-prism refractions, then entering a triple-Laue interferometer. This interferometer was adjusted to 220 reflections. A parallel-sided plate was used as a phase shifter (varying χ). A pair of water-cooled Helmholtz coils produced a fairly uniform magnetic guide field, $B_0 + \hat{z}$, around the interferometer. A magnetically saturated Heusler crystal together with a rectangular spin rotator (adjusting α) and a spin flipper, if necessary, enabled the selection of neutrons with certain polarization directions.

A crucial optical element in our preparation is a spin turner, which turns the incident spinor $|\uparrow\rangle$ to $|\rightarrow\rangle$ in one beam and to $|\leftarrow\rangle$ in the other. For this procedure, we used a soft-magnetic Mu-metal sheet²⁴, which gives high permeability induced by weak magnetic fields. We used a sheet of 0.5 mm thick in an oval ring form and two DC-coils to magnetize this soft-magnetic sheet.

In the manipulation, the important parameters are the relative phase, χ , between the two beams and the spinor rotation angle, α . To show the capability of our manipulation, we measured interference oscillations for two two-level systems in the interferometer: one for a spatial (path), and the other for spinor (spin). Typical oscillations are shown in Fig. 2. Sinusoidal oscillations occur when varying the parameters χ , in Fig. 2a, and α , in Fig. 2b. Sufficiently high contrasts were achieved to confirm the fact that we established an apparatus for manipulating two subsystems: the path and the spin of neutrons in the interferometer.

A maximum violation of the Bell-like inequality is expected in setting the spinor rotation angle α at $0, \pi/2, \pi$ and $3\pi/2$. Typical intensity modulations, obtained by varying the phase shift χ , are shown in Fig. 3. Contrasts evidently decreased from those shown in Fig. 2, mainly because of dephasing/depolarization at the Mu-metal spin-turner. A gradual reduction of contrast by increasing α is attributed to slight depolarization by the spinor-rotator and the

π -spin-flipper. We, however, managed to obtain enough high contrasts, more than 70.7%, to accomplish the experiment. The Mu-metal spin-turner induces additional relative phase shift between two beams in the interferometer, so all interference oscillations are shifted by about π in this figure. We took this shift into account in determining appropriate χ -positions to show the maximum violation.

After fitting to sinusoidal dependence by the least-squares method, the expectation values E_{obs} were calculated using equation (2). The typical statistical error of E_{obs} was ± 0.01 , obtained from curves of a single measurement. We repeated the same measurements at least 16 times to reduce statistical errors. The final value of E_{obs} and its error were evaluated by the weighted average of all measurements. So, the final errors are the sum of systematic and statistical errors. (The main reason for systematic error was phase instability, random drift of phase, during the measurement.) We obtained $E_{\text{obs}}(0, 0.79\pi)$ to be 0.542 ± 0.007 from the intensities of $N'(0, 0.79\pi)$, $N'(\pi, 0.79\pi)$, $N'(0, 1.79\pi)$, and $N'(\pi, 1.79\pi)$. In the same manner, we determined $E_{\text{obs}}(0, 1.29\pi) = 0.4882 \pm 0.012$, $E_{\text{obs}}(0.5\pi, 0.79\pi) = -0.538 \pm 0.006$, and $E_{\text{obs}}(0.5\pi, 1.29\pi) = 0.438 \pm 0.012$. In evaluating the Bell-like inequality, S' was calculated to be $2.051 \pm 0.019 > 2$, for $\alpha_{1,2} = 0, 0.50\pi$, and $\chi_{1,2} = 0.79\pi, 1.29\pi$. This clearly shows a violation of the Bell-like inequality: stronger correlations than the classical non-contextual model.

The results above were obtained using a neutron detector of more than 99% efficiency. In this case, however, a fair-sampling hypothesis is still required, because of losses in the interferometer—the second plate of the interferometer is not a mirror but a beam-splitter—in addition to the fact that the count rates were obtained successively one after another. (It should be mentioned that equation (2) itself does not use this assumption.) A similar experiment concerning the non-contextual hidden variable theories

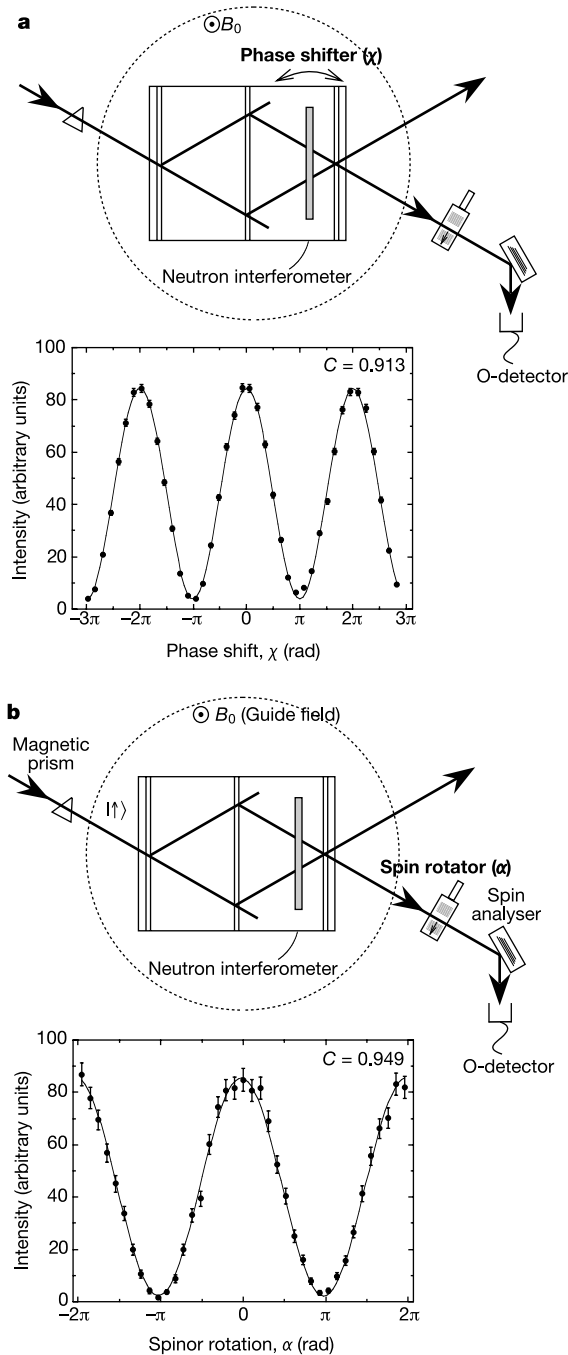


Figure 2 Interference oscillations for two-level systems. **a**, Spatial Hilbert-spaces; **b**, polarization Hilbert-spaces. Contrasts of over 91% for **a** and about 95% for **b** were achieved. This confirmed the quality of our manipulation for the path and spinor properties.

with entangled photons²⁵ was reported. In this experiment, a ‘trigger’ instead of a polarized incident beam was used to show correlations in coincidence counting. This leaves the total system in an entangled state, potentially one of four states with equal probabilities. Thus, the ‘projection’ postulate in the mixed state is needed to deduce the quantum contextuality from this experiment.

The correlations observed in our experiments could be discussed in terms of a semiclassical wave theory, that is, as beam polarizations, as frequently used in optics. This calculation exhibits a non-factorizable total polarization by $P^s(\alpha)$ for the spin manipulations

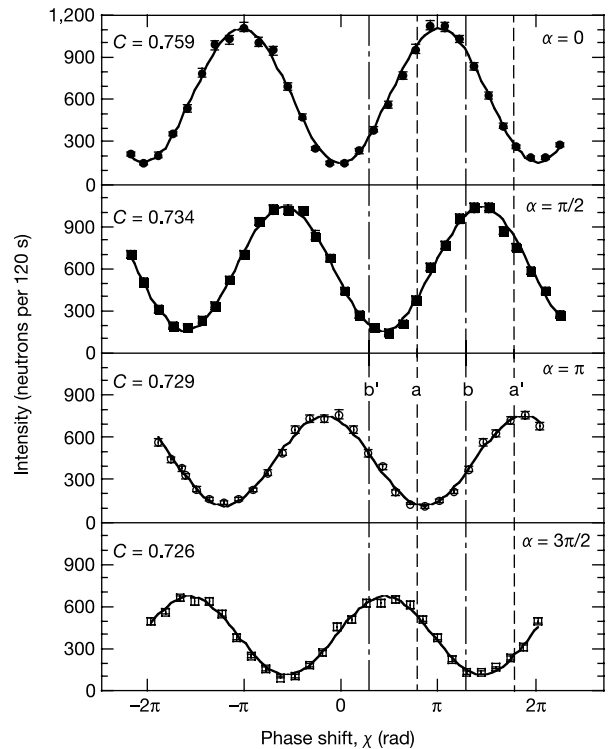


Figure 3 Typical interference oscillations with spinor rotation angle $\alpha = 0, \pi/2, \pi$ and $3\pi/2$. Contrasts of 76% for one and about 73% for the other three were achieved. Expectation values, E_{obs} , were derived from the intensities of appropriate χ : on line a $\chi = 0.79\pi$, on line a' $\chi = 1.79\pi$, on line b $\chi = 1.29\pi$, and on line b' $\chi = 0.29\pi$, where a maximum violation of the Bell-like inequality is expected. These values exhibited the final S' value of $2.051 \pm 0.019 > 2$, which indicates a clear violation of the Bell-like inequality.

and $P^p(\chi)$ for the path manipulations. We note here that the entanglement is not limited to different particles but is generally applicable to different degrees of freedom in single particles. General arguments on the entanglement-induced correlation can be found in the literature^{26,27}.

Received 1 April; accepted 25 June 2003; doi:10.1038/nature01881.

1. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
2. Bell, J. S. On the Einstein-Podolsky-Rosen paradox. *Physics* **1**, 195–200 (1964).
3. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
4. Clauser, J. F. & Shimony, A. Bell's theorem: experiment tests and implications. *Rep. Prog. Phys.* **41**, 1881–1927 (1978).
5. Mann, A. & Revzen, M. (eds) *The Dilemma of Einstein Vol. 12, Podolsky and Rosen: 60 Years Later* (Annals of the Israel Physical Society, The Institute of Physics, Bristol, 1996).
6. Afriat, A. & Selleri, F. (eds) *The Einstein, Podolsky and Rosen Paradox in Atomic, Nuclear and Particle Physics* (Plenum, New York, 1999).
7. Bertlmann, R. A. & Zeilinger, A. (eds) *Quantum [Un]speakeables* (Springer, Berlin/Heidelberg/New York, 2002).
8. Rowe, M. A. *et al.* Experimental violation of a Bell's inequality with sufficient detection. *Nature* **409**, 791–794 (2001).
9. Bell, J. S. On the problem of hidden variables in quantum mechanics. *Rev. Mod. Phys.* **38**, 447–452 (1966).
10. Kochen, S. & Specker, E. P. The problem of hidden variables in quantum mechanics. *J. Math. Mech.* **17**, 59–87 (1967).
11. Mermin, N. D. Simple unified form for the major no-hidden-variables theorems. *Phys. Rev. Lett.* **65**, 3373–3376 (1990).
12. Mermin, N. D. Hidden variables and the two theorems of John Bell. *Rev. Mod. Phys.* **65**, 803–815 (1993).
13. Roy, S. M. & Singh, V. Quantum violation of stochastic noncontextual hidden-variable theories. *Phys. Rev. A* **48**, 3379–3381 (1993).
14. Basu, S., Bandyopadhyay, S., Kar, G. & Home, D. Bell's inequality for single spin-1/2 particle and quantum contextuality. *Phys. Lett A* **279**, 281–286 (2001).
15. Rauch, H. & Werner, S. A. *Neutron Interferometry* (Clarendon, Oxford, 2000).
16. Freedman, S. J. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28**,

- 938–941 (1972).
17. Aspect, A., Grangier, P. & Roger, G. Experimental realization of Einstein-Podolsky-Rosen-Bohm gedanken experiment: A new violation of Bell's inequalities. *Phys. Rev. Lett.* **49**, 91–94 (1982).
 18. Ou, Z. Y. & Mandel, L. Violation of Bell's inequality and classical probability in a two-photon correlation experiment. *Phys. Rev. Lett.* **61**, 50–53 (1988).
 19. Tittel, W., Brendel, J., Zbinden, H. & Gisin, N. Violation of Bell inequalities by photons more than 10 km apart. *Phys. Rev. Lett.* **81**, 3563–3566 (1998).
 20. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
 21. Sakurai, J. J. *Modern Quantum Mechanics* 203 (Addison-Wesley, Reading, Massachusetts, 1994).
 22. Aspect, A. in *The Wave-Particle Dualism* (eds Diner, S., Farque, D., Lochak, G. & Selleri, F.) 377 (D. Reidel Publishing Company, Dordrecht, 1984).
 23. Kroupa, G. *et al.* Basic features of the upgraded S18 neutron interferometer set-up at ILL. *Nucl. Instrum. Meth. A* **440**, 604–608 (2000).
 24. Rauch, H., Wilfing, A., Bauspiess, W. & Bonse, U. Precise determination of the 4 π -periodicity factor of a spinor wave function. *Z. Phys. B* **29**, 281–284 (1978).
 25. Michler, M., Weinfurter, H. & Zukowski, M. Experiments towards falsification of noncontextual hidden variable theories. *Phys. Rev. Lett.* **84**, 5457–5461 (2000).
 26. Englert, B.-G. Remark on some basic issues in quantum mechanics. *Z. Naturforsch.* **54a**, 11–32 (1999).
 27. Bell, J. S. in *Themes in Contemporary Physics II* (eds Deser, S. & Finkelstein, R. J.) 1 (World Scientific, Singapore, 1989).

Acknowledgements We appreciate discussions with R. A. Bertlmann, I. J. Cirac and D. Home. This work has been supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.H. (hasegawa@ati.ac.at).

Entangled quantum state of magnetic dipoles

S. Ghosh¹, T. F. Rosenbaum¹, G. Aeppli² & S. N. Coppersmith³

¹James Franck Institute and Department of Physics, University of Chicago, Chicago, Illinois 60637, USA

²London Centre for Nanotechnology and Department of Physics and Astronomy, University College London, London WC1E 6BT, UK, and NEC Laboratories, 4 Independence Way, Princeton, New Jersey 08540, USA

³Department of Physics, University of Wisconsin, Madison, Wisconsin 53706, USA

Free magnetic moments usually manifest themselves in Curie laws, where weak external magnetic fields produce magnetizations that vary as the reciprocal of the temperature ($1/T$). For a variety of materials that do not display static magnetism, including doped semiconductors¹ and certain rare-earth intermetallics², the $1/T$ law is replaced by a power law $T^{-\alpha}$ with $\alpha < 1$. Here we show that a much simpler material system—namely, the insulating magnetic salt $\text{LiHo}_x\text{Y}_{1-x}\text{F}_4$ —can also display such a power law. Moreover, by comparing the results of numerical simulations of this system with susceptibility and specific-heat data³, we show that both energy-level splitting and quantum entanglement are crucial to describing its behaviour. The second of these quantum mechanical effects—entanglement, where the wavefunction of a system with several degrees of freedom cannot be written as a product of wavefunctions for each degree of freedom—becomes visible for remarkably small tunnelling terms, and is activated well before tunnelling has visible effects on the spectrum. This finding is significant because it shows that entanglement, rather than energy-level redistribution, can underlie the magnetic behaviour of a simple insulating quantum spin system.

The insulator that we focus on in the search for the cause of the anomalous power-law divergence of the magnetic susceptibility is $\text{LiHo}_x\text{Y}_{1-x}\text{F}_4$, a salt where magnetic Ho^{3+} ions are randomly

substituted for nonmagnetic Y^{3+} with probability x . For $x = 1$, the material is the dipolar-coupled ferromagnet, LiHoF_4 , with a Curie temperature of 1.53 K. Randomly distributing dipoles in a solid matrix provides quenched disorder, while the angular anisotropy of the dipole–dipole interaction leads to competition between ferromagnetic and antiferromagnetic bonds and the possibility of many (nearly) degenerate ground states⁴. Indeed, the low-temperature magnetic phase diagram of the dipolar-coupled rare-earth tetrafluorides progresses smoothly from long-range order to glassiness with increasing spin dilution³. What interests us here, however, is the considerably diluted $x = 0.045$ compound, where we have observed^{5,6}—contrary to classical expectations⁴—novel ‘antiglass’ behaviour as well as long-lived spin oscillations whose qualitative understanding seems to require mesoscopic quantum coherence. We show in Fig. 1 the experimental d.c. susceptibility, χ , plotted against temperature, T , for a single-crystal specimen of the material. What emerges is not the standard Curie law, $1/T$, expected for non-interacting magnetic moments, but instead a diverging response following a power law $T^{-\alpha}$, with $\alpha = 0.75 \pm 0.01$. This power law is close to that associated with the diverging local susceptibilities inferred for doped silicon¹ as well as metallic rare-earth materials² on the brink of magnetic order. What is most striking, however, is that the magnetic susceptibility for $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$ is a smoothly diverging quantity, even though the magnetic specific heat (C , Fig. 2a) is characterized by unusually sharp peaks in the same temperature range. In ordinary materials containing magnetic ions, there is a strong correlation between magnetic susceptibility and specific heat in the sense that anomalies, especially as strong as the

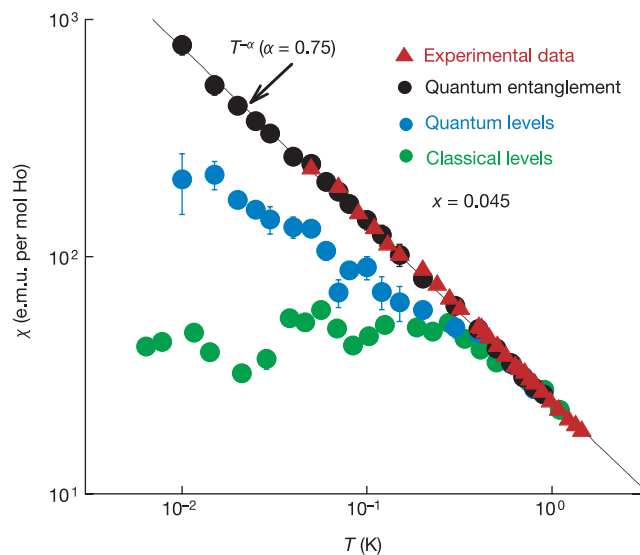


Figure 1 Magnetic susceptibility χ versus temperature T of the diluted, dipolar-coupled Ising magnet, $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$. Red triangles, experimental data; filled circles, simulations. Green circles, classical decimation when the calculations are performed with $g_{\perp} = 0$. Blue circles, susceptibility computed using the classical procedure, equation (3), of determining Curie constants by adding (subtracting) moments when the ground state is predominantly ferromagnetic (antiferromagnetic), but with quantum decimation, using energy levels derived from the full dipolar hamiltonian of equation (1). Although the susceptibility approaches that of the experiment more closely than before, it still deviates by at least a factor of four at low temperatures. Black circles, use of quantum decimation as well as the correct quantum mechanical form of susceptibility given by equation (5), using the entanglement of the low-lying energy doublet with the excited states. The line is a best fit to $\chi(T) \propto T^{-\alpha}$, with $\alpha = 0.75 \pm 0.01$. Although α is always less than 1, it is not a universal number. It varies from 0.62 to 0.81 as the concentration x decreases from 0.1 to 0.01, a trend also observed in Heisenberg systems⁷. The simulation results have not been scaled, and agree quantitatively with the experimental results.

sharp peaks in the specific heat, are reflected in the susceptibility.

The data for $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$ thus provide three puzzles: the absence of a spin glass transition predicted for a collection of randomly placed dipoles, the anomalous power-law behaviour $\chi \propto T^{-\alpha}$, and the coexistence of a featureless power law in χ with sharp anomalies in the specific heat. We show here that it is an intrinsic quantum mechanical term in the dipole hamiltonian that stabilizes the spin liquid 'antiglass' and resolves the puzzles. Following a pair-wise 'decimation' procedure^{7–13} adapted to treat the full axial and transverse components of the dipole–dipole interaction, we find that quantum fluctuations continue to provide channels for relaxation down to the lowest temperatures. We simulate the evolution with temperature T of both the magnetic susceptibility and the heat capacity using the actual interaction parameters between Ho moments obtained from various experimental results, and compare quantitatively the results of simulation and experiment.

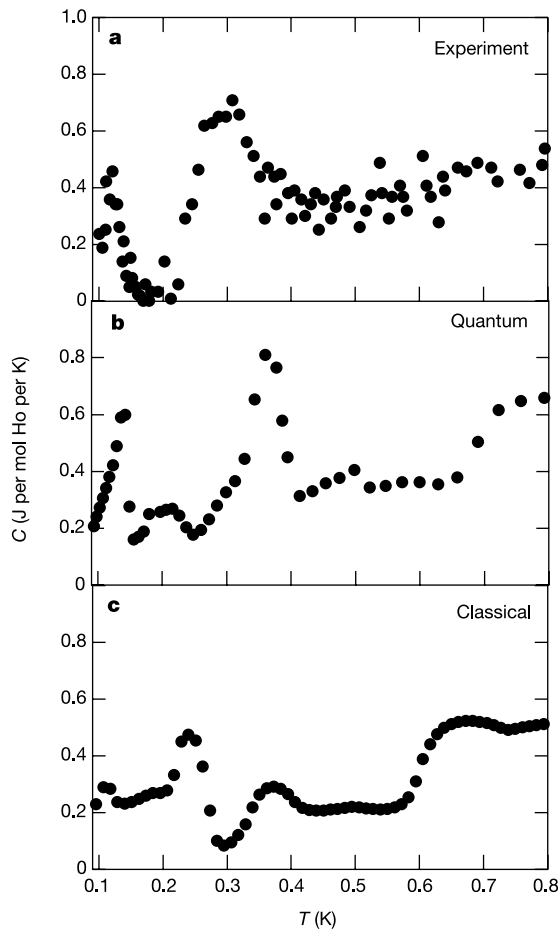


Figure 2 Comparison of the temperature-dependent experimental electronic specific heat $C(T)$ for $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$ with different simulation techniques. **a**, Experimental data showing two sharp Schottky-like features dominating the thermal response. **b**, Quantum decimation, emphasizing how the well-defined energy levels result in a more complex temperature-dependent specific heat with greater resemblance to the experimental data, especially at low temperatures. Notably, there is the appearance of a sharp peak at 130 mK, close to a similar feature in the data. **c**, Classical decimation, demonstrating some success in calculating $C(T)$ but with the characteristic sharp features occurring at incorrect temperatures. The features occur as $k_B T$ moves through maxima in the distribution of dipolar couplings, and are affected as the concentration x varies; this distribution is granular, because the dipolar interaction is being sampled between points on a lattice rather than in continuous space.

LiHoF_4 crystallizes in a body-centred tetragonal (CaWO_4) structure with lattice constants¹⁴ $a = a' = 5.175(5) \text{ \AA}$ and $c = 10.75(1) \text{ \AA}$. Each unit cell has four formula units with the magnetic Ho^{3+} ions occupying positions $(0, 0, 0)$, $(0, a/2, c/4)$, $(a/2, a/2, c/2)$ and $(a/2, 0, 3c/4)$. The Ising axis is defined by the crystal field of the Ho^{3+} ions that forces the spin-1/2 magnetic moments with a g -factor of 13.8 to point along the crystalline c axis. We generate a model of the three-dimensional lattice of $\text{LiHo}_x\text{Y}_{1-x}\text{F}_4$ on the computer by repeated translation of the unit cell vectors. N spins are distributed randomly in this lattice with probability of occupancy x on the body-centred tetragonal sites. Periodic boundary conditions are applied. We have used a maximum of $N = 400$ spins (8×10^4 pair-wise interactions), and have checked that our results are in the N -independent limit. The Ising spin (σ_i^z) at each site i is assigned a value of 1 or -1 randomly. We have confirmed explicitly that in the dilute limit of x the outcome is not sensitive to this particular initial condition, obviating the need to average over initial spin configurations. Our simulations incorporate pair-wise dipolar couplings, in which the interaction energy between two magnetic dipole moments $\mathbf{M}_1$ and $\mathbf{M}_2$ separated by the vector $\mathbf{r} = \hat{\mathbf{r}}r$ is

$$E_{\text{int}} = \frac{1}{r^3} (\mathbf{M}_1 \cdot \mathbf{M}_2 - 3(\hat{\mathbf{r}} \cdot \mathbf{M}_1)(\hat{\mathbf{r}} \cdot \mathbf{M}_2)) \quad (1)$$

The magnetic dipole moment $\mathbf{M}_i$ at site i is related to the spin σ_i via $M_{i\lambda} = \mu_B g_{i\lambda} \sigma_{i\lambda}$ ($\lambda = x, y, z$) where μ_B is the magnetic moment of the i th spin, initially $\mu_i = \mu_B/2$ for all i , and the elements of the anisotropic g -factor matrix ($g_x = g_y = g_{\perp} = 0.74$, $g_z = g_{\parallel} = 13.8$) are known from previous measurements on the pure material^{14–16}.

Our first simulation is a classical calculation in which $g_{\perp} = 0$ and equation (1) reduces to $-2M_{1z}M_{2z}/r^3$. The hamiltonian can be

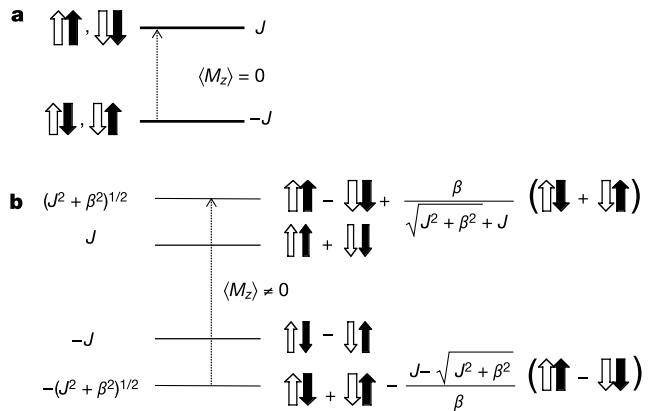


Figure 3 Diagram detailing the difference between the classical and the quantum decimation schemes. **a**, The classical energy levels calculated using equation (2). There are two doubly degenerate energy levels, designated $+J$ and $-J$. Depending on the value of the angle formed by the Ising axis and the vector connecting the spins, the ground state $-J$ can correspond to antiferromagnetic or ferromagnetic alignment of (i, j) . The eigenstates commute with σ_z , and there is no mixing between the ground-state doublet and the excited states. **b**, The quantum energy levels, showing the new entangled eigenstates. β is the off-diagonal term of the full dipolar interaction of equation (1), which reduces to the well-known hamiltonian of Ising spins in a transverse magnetic field in the limit of small $(g_{\perp}/g_{\parallel})^2$. The two doublets are split to produce eigenstates that are mixtures of the classical states. It is not only states from the same doublet that are mixed: β also yields a ground state that mixes ferromagnetic and antiferromagnetic classical states. Thus, the off-diagonal terms in the dipolar interaction introduce both a change in the spectrum (in the form of splittings of the doublets as well as shifts of the 'ferromagnetic' excited states relative to the 'antiferromagnetic' ground state), and mixing (or 'entanglement') of the classical states.

written as:

$$H = - \sum_{ij} J_{ij} \mu_i \sigma_{iz} \mu_j \sigma_{jz} \quad (2)$$

where J_{ij} (which assimilates the numerical constants and the product $g_{\parallel}^2$) falls off as $1/r^3$ and the quantities denoted by σ are classical Ising spins that can take the values ± 1 . It is expected to be valid both in the very dense ($x = 1$) ferromagnetic and very dilute ($x \rightarrow 0$) paramagnetic limits. We call equation (2) an ‘axial dipole’ hamiltonian, as the dipolar field of spin j coupling to the moment of spin i acts only along the Ising axis. We show in Fig. 3a the energy levels calculated using equation (2) for a spin pair (i, j) when $|\mu_i| = |\mu_j|$.

Once we have calculated the energy of all spin pairs, we arrange the pairs in a hierarchy based on their coupling strength and pick the pair with the largest energy $|J_{\max}|$. If $2|J_{\max}| > k_B T$, the excited state $+J_{\max}$ becomes redundant and this pair is forced into its ground state. The pair is replaced by a composite single spin of equivalent net magnetic moment that can be either $\mu_C = |\mu_i| - |\mu_j|$ (anti-ferromagnetic interaction) or $\mu_C = |\mu_i| + |\mu_j|$ (ferromagnetic interaction). If the net moment is zero, the two spins are ‘decimated’ (that is, eliminated) and removed from further consideration; otherwise they are replaced by one spin with the new composite moment placed at the average position of the two spins in the pair. Only the magnetic moment μ_C and the position $\mathbf{r}_C$ of the pair are renormalized; $g_{\parallel}$ is left unchanged. The new magnetic dipole moment M_{Cz} is now given by $\mu_C g_{\parallel} \sigma_{Cz}$ ($M_{Cx} = M_{Cy} = 0$). The altered demography requires the procedure to begin anew with the calculations of the pair-wise interactions between the remaining spins and the composite $\mathbf{M}_C$ at $\mathbf{r}_C$ using equation (2). At each iteration one spin pair is either eliminated or transformed into a single spin degree of freedom until the strongest pair remaining has a gap $2|J_{\max}| < k_B T$. All spins remaining in the system are considered ‘free’ at that T . We choose $T = 1$ K at the outset, given that

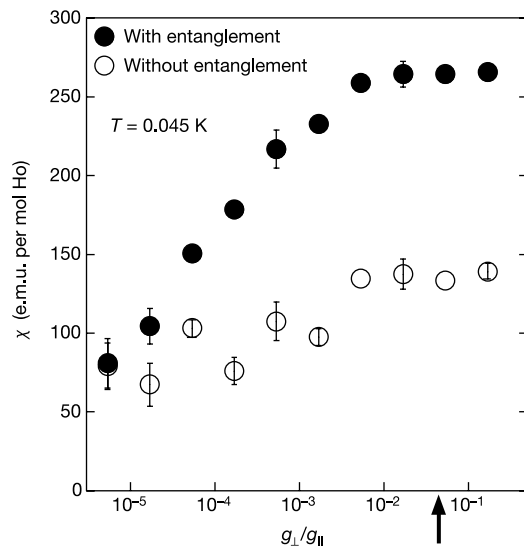


Figure 4 The change in susceptibility as the quantum entanglement is tuned by varying the ratio of the transverse and longitudinal magnetic g -factors, $g_{\perp}/g_{\parallel}$. An arrow denotes the value for $\text{LiHo}_{0.45}\text{Y}_{0.55}\text{F}_4$. The full quantum susceptibility (filled circles) demonstrates extraordinary sensitivity to the slightest entanglement of the wavefunctions ($g_{\perp}/g_{\parallel} \approx 10^{-4}$), whereas the susceptibility calculated using the quantum energy levels but ignoring the entanglement (open circles) is relatively flat. Quantum entanglement produced by the off-diagonal terms, rather than spectral superposition, dominates the physics.

the nearest-neighbour interaction energy $J_{nn} = 1.2$ K; the temperature is then lowered in steps $\Delta T = 0.01$ K down to $T = 0.01$ K.

At each temperature, the calculation produces a list $\{M_i\}$ of $N(T)$ ‘free’ moments. Given that the susceptibility for a free Ising moment M_i is $M_i^2/k_B T$ and its contribution to the entropy is $(R \ln 2)N(T)$, we can compute the specific heat and susceptibility via the relations:

$$\chi = \sum_i \frac{M_i^2}{k_B T} \quad (3)$$

and

$$C = \frac{T dS}{dT} = RT \ln 2 \frac{dN(T)}{dT} \quad (4)$$

Figure 2c reveals one success of the classical decimation approach, namely the appearance of sharp features in the specific heat. Although the features are of roughly the same magnitude as seen in the experiment, they do not occur at the correct temperatures. Figure 1, where the filled green circles are the results of the classical calculation, reveals a more disturbing problem. In accord with intuition, but in disagreement with experiment, there are sharp anomalies in χ that coincide with the peaks in C . Moreover, the classical susceptibility at low temperatures is over an order of magnitude smaller than that measured.

The axial dipole hamiltonian in equation (2) takes into account only the interactions parallel to the Ising axis by ignoring $g_{\perp}$ and treats the spins as classical bits rather than Pauli matrices. For pure LiHoF_4 in its ferromagnetic state, this is the complete picture, because the lattice symmetry ensures that at any site the perpendicular component of the dipolar field due to the other spins sums to zero. However, as the magnetic Ho^{3+} ions are randomly replaced by non-magnetic Y^{3+} , all transverse components are no longer perfectly compensated, and at dipole concentrations of only a few per cent the site-specific, internal transverse fields can be as large as 1 kOe. With a finite $g_{\perp}$, the full hamiltonian, equation (1), no longer reduces to equation (2), but acquires off-diagonal terms of the form $\beta \sigma_{ix} \sigma_{jz}$, where β includes numerical constants and the product $g_{\perp} g_{\parallel}$. Terms of order $\sigma_{ix} \sigma_{jx}$ are not included, because $g_{\perp}^2 \ll g_{\perp} g_{\parallel} \ll g_{\parallel}^2$.

The results in Figs 1 and 2 demonstrate how the decimation calculation is affected when the energy levels but not the eigenfunctions are modified to account for the inclusion of the off-diagonal terms of the dipolar interaction. Although the basic decimation scheme remains the same, there are now two energy scales available for comparison with temperature: $2|J|$ and $(J^2 + \beta^2)^{1/2} - J$, which is $\sim \beta^2/2J$ to first order. As this second gap is much smaller than the first, it becomes clear that, at a temperature T , the number of free spins, $N(T)$, is greater than in the classical case (see Fig. 3b). The results of this modification meet with partial success. The increase in $N(T)$ can account better for the specific-heat characteristics (Fig. 2b), but not for the susceptibility, especially at low temperatures (Fig. 1, filled green circles), implying that it is not merely the excess in the number of free spins that enhances and smoothes the susceptibility of the sample.

The key to matching the experimental susceptibility result is to use the quantum mechanical expression¹⁷ for χ :

$$\chi = \sum_i \frac{1}{\sum_n \rho_n(i)} \sum_n \left(\frac{(E_n^{(1)}(i))^2}{k_B T} - 2E_n^{(2)}(i) \right) \rho_n(i) \quad (5)$$

with

$$\rho_n(i) = \exp\left(\frac{-E_n^{(0)}(i)}{k_B T}\right); \quad E_n^{(1)}(i) = \langle n(i) | M_{iz} | n(i) \rangle \quad \text{and}$$

$$E_n^{(2)}(i) = \sum_m \frac{|\langle n(i) | M_{iz} | m(i) \rangle|^2}{E_m^{(0)}(i) - E_n^{(0)}(i)}$$

where first we sum over the energy levels of the i th effective spin remaining, and then sum over all N effective spins. Except for a population of isolated, unrenormalized spins that dwindles as T decreases, $|n\rangle$ and $|m\rangle$ are now the entangled pair eigenstates illustrated in Fig. 3b. Note that keeping only the first term in the brackets reduces equation (5) to equation (3). The result obtained using equation (5) (Fig. 1, black circles) agrees quantitatively with the actual measurements (Fig. 1, red triangles) of the magnetic susceptibility in the d.c. limit for $\text{LiHo}_{0.045}\text{Y}_{0.955}\text{F}_4$.

The inclusion of the off-diagonal terms, $E_n^{(2)}$, produces the reconciliation between simulation and experiment by superposing the antiferromagnetic ground state with the ferromagnetic excited state (Fig. 3b). In this manner, the excited classical states 'frozen out' by the decimation still enter into the expression for the susceptibility, thereby enhancing its value. The concurrence^{18,19}, which ranges from 0 (disentangled states) to 1 (completely entangled), quantifies entanglement in an exact way for bipartite systems. We identify the T -dependent entanglement τ as the concurrence of the pair wavefunctions contributing to the susceptibility at each T , weighted by the fraction of actual spins involved in the history of the pair formation within the decimation calculation. We find that $\tau = 0.11$ at 0.8 K, growing to 0.88 at 0.01 K, in accord with the trend from near agreement of the 'entangled' and 'quantum level' bulk susceptibilities at 0.8 K towards a factor of four difference at 0.01 K. Moreover, we find that only the slightest degree of entanglement can have profound effects. We illustrate this in Fig. 4, which shows how the two calculations evolve with increasing $g_{\perp}$. The figure reveals that a $g_{\perp}/g_{\parallel}$ as small as 10^{-4} produces a large change in χ . The effects of the energy-level distribution are minimal by comparison.

Contrary to intuition and common experience, a dilute assembly of Ising dipoles does not freeze when cooled to millikelvin temperatures. The computer simulations presented here indicate that it is quantum mechanics—the internal magnetic fields transverse to the Ising axis inherent to the dipole–dipole interaction—that stabilizes the spin liquid. However, unlike conventional spin liquids, where the dynamics are dominated by a single gap to triplet excitations, the dilute dipoles form a state with a distribution of such gaps, especially well probed by the specific heat, which shows remarkable releases of entropy at certain well-defined temperatures. At the same time, the magnetic susceptibility increases smoothly with decreasing temperature, but at a rate slower than a Curie law⁶. The smoothness is in marked contrast to the highly structured heat capacity, and can only be understood if quantum mechanical mixing—the entanglement of classical ferromagnetic and antiferromagnetic contributions to the wavefunctions—is taken into account. There is a growing realization^{20–22} that entanglement is a useful concept for understanding quantum magnets, thus unifying two rapidly evolving areas, quantum information theory and quantum magnetism. The discussions to date have focused on one-dimensional magnets and measures of entanglement with clear theoretical meaning but no simple experimental implementation. Our experiments and simulations illustrate how entanglement, rather than energy-level redistribution, can contribute significantly to the simplest of observables—the bulk susceptibility—in an easily stated model problem. □

Received 29 April; accepted 7 July 2003; doi:10.1038/nature01888.

1. Paalanen, M. A., Ruckenstein, A. E. & Thomas, G. A. Spins in Si:P close to the metal–insulator transition. *Phys. Rev. Lett.* **54**, 1295–1298 (1985).
2. Schröder, A. *et al.* Onset of antiferromagnetism in heavy-fermion metals. *Nature* **407**, 351–355 (2000).
3. Reich, D. H. *et al.* Dipolar magnets and glasses: Neutron scattering, dynamical, and calorimetric studies of randomly distributed Ising spins. *Phys. Rev. B* **42**, 4631–4644 (1990).
4. Aharony, A. & Stephen, M. J. Percolation with long-range interactions. *J. Phys. C* **14**, 1665–1670 (1981).
5. Reich, D. H., Rosenbaum, T. F. & Aeppli, G. Glassy relaxation without freezing in a random dipolar-coupled Ising magnet. *Phys. Rev. Lett.* **59**, 1969–1972 (1987).
6. Ghosh, S., Parthasarathy, R., Rosenbaum, T. F. & Aeppli, G. Coherent spin oscillations in a disordered magnet. *Science* **296**, 2195–2198 (2002).
7. Bhatt, R. N. & Lee, P. A. Scaling studies of highly disordered spin-1/2 antiferromagnetic systems. *Phys. Rev. Lett.* **48**, 344–347 (1982).

8. Ma, S. K., Dasgupta, C. & Hu, C. K. Random antiferromagnetic chain. *Phys. Rev. Lett.* **43**, 1434–1437 (1979).
9. Dasgupta, C. & Ma, S. K. Low temperature properties of the random Heisenberg antiferromagnetic chain. *Phys. Rev. B* **22**, 1305–1319 (1980).
10. Fisher, D. S. Random transverse field Ising spin chains. *Phys. Rev. Lett.* **69**, 534–537 (1992).
11. Fisher, D. S. Critical behaviour of random transverse-field Ising spin chains. *Phys. Rev. B* **51**, 6411–6461 (1995).
12. Westerberg, E., Furusaki, A., Sigrist, M. & Lee, P. A. Random quantum spin chains: A real space renormalization group study. *Phys. Rev. Lett.* **75**, 4302–4305 (1995).
13. Phillips, P., Izzo, D. & Kundu, K. Spin-lattice relaxation below 1 K: A new mechanism for unexpected nuclear spin relaxation. *Phys. Rev. B* **37**, 10876–10879 (1988).
14. Hansen, P. E., Johansson, T. & Nevald, R. Magnetic properties of lithium rare-earth fluorides: Ferromagnetism in LiErF_4 and LiHoF_4 and crystal-field parameters at the rare-earth and Li sites. *Phys. Rev. B* **12**, 5315–5324 (1975).
15. Bitko, D., Rosenbaum, T. F. & Aeppli, G. Quantum critical behaviour for a model magnet. *Phys. Rev. Lett.* **77**, 940–943 (1996).
16. Magarino, J., Tuchendler, J., Beauvillain, P. & Laursen, I. EPR experiment in LiTbF_4 , LiHoF_4 , and LiErF_4 at submillimeter frequencies. *Phys. Rev. B* **21**, 18–28 (1980).
17. Van Vleck, J. H. Quantum mechanics—the key to understanding magnetism. *Rev. Mod. Phys.* **50**, 181–189 (1978).
18. Gunlycke, D., Kendon, V. M., Vedral, V. & Bose, S. Thermal concurrence mixing in a one-dimensional Ising model. *Phys. Rev. A* **64**, 042302 (2001).
19. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80**, 2245–2248 (1998).
20. Osterloh, A., Amico, L., Falci, G. & Fazio, R. Scaling of entanglement close to a quantum phase transition. *Nature* **416**, 608–610 (2002).
21. Arnesen, M. C., Bose, S. & Vedral, V. Natural thermal and magnetic entanglement in the 1D Heisenberg model. *Phys. Rev. Lett.* **87**, 017901 (2001).
22. Osborne, T. J. & Nielsen, M. A. Entanglement, quantum phase transitions, and density matrix renormalization. *Quant. Inform. Process.* **1**, 45–53 (2002).

Acknowledgements We thank R. Parthasarathy for discussions. The work at the University of Chicago was supported by the MRSEC Program of the National Science Foundation, that in the University of Wisconsin by the Petroleum Research Fund and the National Science Foundation, and that in University College London by a Wolfson–Royal Society Research Merit Award and the Basic Technologies programme of the UK Research Councils.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.F.R. (t-rosenbaum@uchicago.edu).

Magnetic enhancement of superconductivity from electron spin domains

H. A. Radovan¹, N. A. Fortune², T. P. Murphy¹, S. T. Hannahs¹, E. C. Palm¹, S. W. Tozer¹ & D. Hall^{1*}

¹NHML, Florida State University, Tallahassee, Florida 32310, USA

²Department of Physics, Smith College, Northampton, Massachusetts 01063, USA

* Present address: American Physical Society, One Research Road, Box 9000, Ridge, New York 11961, USA

Since the discovery of superconductivity¹, there has been a drive to understand the mechanisms by which it occurs. The BCS (Bardeen–Cooper–Schrieffer) model successfully treats the electrons in conventional superconductors as pairs coupled by phonons (vibrational modes of oscillation) moving through the material², but there is as yet no accepted model for high-transition-temperature, organic or 'heavy fermion' superconductivity. Experiments that reveal unusual properties of those superconductors could therefore point the way to a deeper understanding of the underlying physics. In particular, the response of a material to a magnetic field can be revealing, because this usually reduces or quenches superconductivity. Here we report measurements of the heat capacity and magnetization that show that, for particular orientations of an external

magnetic field, superconductivity in the heavy-fermion material CeCoIn₅ is enhanced through the magnetic moments (spins) of individual electrons. This enhancement occurs by fundamentally altering how the superconducting state forms, resulting in regions of superconductivity alternating with walls of spin-polarized unpaired electrons; this configuration lowers the free energy and allows superconductivity to remain stable. The large magnetic susceptibility of this material leads to an unusually strong coupling of the field to the electron spins, which dominates over the coupling to the electron orbits.

In the standard BCS model, orbital effects limit superconductivity, which is destroyed when the kinetic energy of the charges exceeds the condensation energy of the Cooper pairs. The expected zero temperature critical field is given by Werthamer–Helfand–Hohenberg formula $B_{\text{orb}}(T=0) = 0.7T_c|dB_{c2}/dT|_{T_c}$, where T_c is the superconducting transition temperature at zero field and B_{c2} is the upper critical field³. In the paramagnetic or Pauli limit, superconductivity is destroyed when the polarization energy of the spins exceeds the condensation energy due to partial alignment of the spins in field. Here, the zero temperature critical field is given by the Clogston limit, $B_p(T=0) = \Delta_0/\sqrt{2}\mu_B = 1.85T_c$ (where T_c is expressed in kelvin), where Δ_0 is the energy gap, and μ_B is the Bohr magneton⁴. The ratio of these limits, $\alpha = \sqrt{2}B_{\text{orb}}/B_p$, is known as the Maki parameter, and indicates whether a material is orbitally or paramagnetically limited⁵.

In 1964 Fulde and Ferrell⁶ and, independently, Larkin and Ovchinnikov⁷, predicted an additional phase transition within the superconducting state, where the superconducting order parameter

(Cooper pair condensate) lowers its free energy by becoming spatially inhomogeneous. This ‘FFLO’ phase, which consists of a periodically modulated order parameter and an array of Bloch-type magnetic walls of broken Cooper pairs with polarized spins, is an example of spontaneous symmetry breaking. The Bloch wall separation is field dependent, but is of the order of the coherence length, ξ . Within the BCS picture, this state corresponds to Cooper pairs having a finite centre-of-mass momentum due to coupling across the Zeeman-energy-split Fermi surface for up- and down-spin electrons. It is thought that the FFLO transition occurs at or above B_p , thereby raising the superconducting-to-normal transition to higher fields. The temperature of emergence of the FFLO state, T_{FFLO} , cannot exceed $0.56T_c$, but the influence of material parameters may result in lower values^{8,9}.

The FFLO scenario was generalized in 1996 by taking into account both orbital and paramagnetic effects^{10–13}. An experimental realization, in this case, would be a quasi-two-dimensional superconductor in an applied magnetic field slightly tilted away from the conducting planes. The field can be resolved into components normal and parallel to the conducting planes with the components being subject to orbital or paramagnetic limits, respectively. With increasing magnetic field and at a fixed temperature, $T < T_{\text{FFLO}}$, a cascade of first order transitions is expected to occur within the FFLO wedge before eventually reaching the upper critical field, H_{c2} . Each transition indicates an increase in orbital momentum of the superconducting order parameter and corresponds to a specific Landau level state composed of multiquanta vortices. The angle between the external field and the superconducting planes deter-

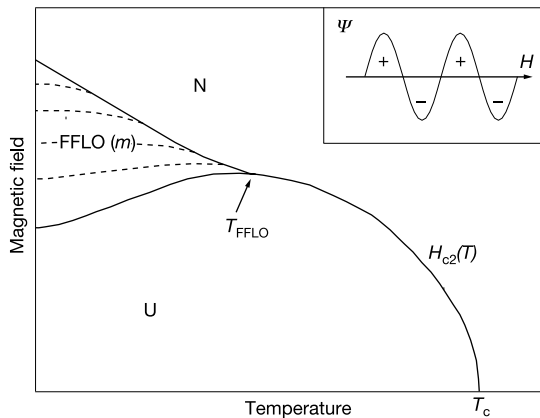


Figure 1 Qualitative phase diagram. Depicted are the uniform superconducting (U), normal (N) and FFLO phases. The inset shows the spatial modulation of the order parameter, Ψ , along the applied field, H , within the non-uniform FFLO phase. Spin-aligned Bloch walls form in planes perpendicular to the magnetic field at the locations of the nodes of the order parameter. The exact temperature of emergence, T_{FFLO} , and the temperature dependence of the uniform superconducting-to-FFLO state transition line depend on the microscopic parameters of the material, such as Fermi surface geometry and spin exchange interactions²⁶. Dashed lines within the FFLO phase illustrate higher Landau level states of the spatially varying order parameter with orbital momentum, m . The solution of the linearized gap equation corresponds to the Schrödinger equation with a field term^{10,11}. The resulting eigenfunctions for the order parameter are $\Delta(r) \approx \exp(-im\varphi)\exp(-\rho^2/2)\rho^m\exp(iQz)$, with orbital quantum number, m , polar angle, φ , dimensionless coordinate, $\rho = (r/2\pi\hbar/hc)^{1/2}$, and magnitude of the FFLO wave vector Q directed along the applied field H in the z direction. This general solution is a set of Landau level functions possessing quantized orbital momentum, m , and $2\pi m$ phase change around $r = 0$. The spacing of these transitions is not simply periodic in $1/B$ as expected for general quantum oscillations, but rather depends on the energies of different Landau level vortex states, a problem that has not been solved. The flux line lattice is the superposition of all solutions, corresponding to multiquanta vortex states instead of the usual Abrikosov vortex lattice.

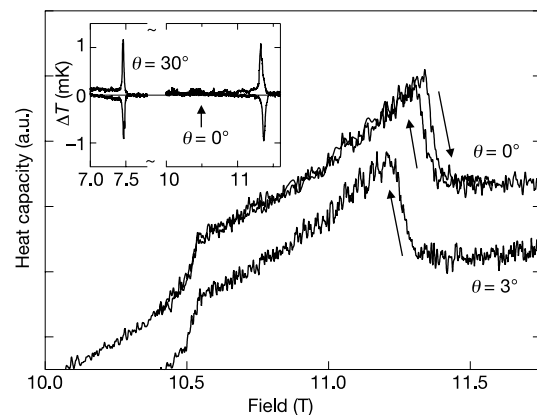


Figure 2 Heat capacity and magnetocaloric measurements. Shown are data as function of field at $T = 250$ mK for CeCoIn₅. We used a custom rotatable calorimeter placed in the mixing chamber of a top loading dilution refrigerator²⁷. The temperature was held constant by actively compensating for the magnetoresistance of the temperature sensors while continuously sweeping field at 0.003 T min^{−1} (ref. 28). Heat capacity for increasing and decreasing field is shown at $\theta = 0^\circ$ and for decreasing field at $\theta = 3^\circ$. The 3° down sweep has been vertically offset and the 3° up sweep has been omitted for clarity. Arrows indicate the direction of the field sweeps. At small angles, when the magnetic field is near the plane parallel orientation ($\theta = 0^\circ$), two phase transitions are present: a lower-field second-order phase transition between the uniform and FFLO superconducting states plus a higher-field first-order phase transition at H_{c2} between the FFLO and normal states. The lower-field FFLO transition is only present for $\theta \leq 10^\circ$. The inset shows magnetocaloric measurements at the same temperature and a higher sweep rate of 0.03 T min^{−1} for $\theta = 0^\circ$ and $\theta = 3^\circ$. These field sweeps show the absorption and release of latent heat at H_{c2} for increasing and decreasing fields, respectively, confirming the first-order nature of the transition at H_{c2} . The absence of any latent heat for the $\theta = 0^\circ$ data at the position marked by the arrow confirms that the FFLO transition is second order. The H_{c2} phase transition for larger angles is still first order, as can be seen from the $\theta = 30^\circ$ data.

mines the highest Landau level orbital quantum number, m , and hence the number of crossed phase transition lines, H_m . The physical FFLO-to-normal state boundary, $H_{c2}(T)$, is the envelope of all $H_m(T)$ lines. Schematically shown in Fig. 1 are the upper critical field, $H_{c2}(T)$, the emerging spatially varying FFLO phase at low temperatures and high magnetic fields, and a series of dashed lines within the FFLO state denoting sub-phases with higher orbital momentum. Theory predicts a jump in the magnetization or critical current when entering each sub-phase m (refs 10, 11). This scenario should be realized for a Maki parameter, $\alpha > 9$ (ref. 10). To summarize, in the pure orbital limit, the energetically favourable solution is the lowest Landau level ($m = 0$). In the pure paramagnetic case, the emerging state is the FFLO phase, while a mixture of both yields new exotic states with centre-of-mass and orbital momenta.

Stringent requirements are imposed on the material for it to be within the paramagnetic limit: a low T_c ; a high effective electron mass, m^* , indicated by a high critical field slope at T_c ; and a mean free path, l , of the electrons that far exceeds the coherence length, $l \gg \xi$, putting the material in the clean limit. In addition, materials possessing a high magnetic susceptibility will tend to be paramagnetically limited. Two candidate classes of materials are the heavy fermion and low-dimensional molecular superconductors. Our measurements were performed on single crystals of the heavy fermion superconductor, CeCoIn₅, with $T_c = 2.3$ K (ref. 14), an initial in-plane field slope of $|dH_{c2}/dT|_{T_c} = 24$ T K⁻¹ (ref. 15), and $l/\xi = 14$ (ref. 16). With a Maki parameter of $\alpha \equiv \sqrt{2B_{orb}/B_p}$, approximately 13, CeCoIn₅ is clearly in the paramagnetic limit, making possible the observation of the FFLO state and the cascade of first order transitions.

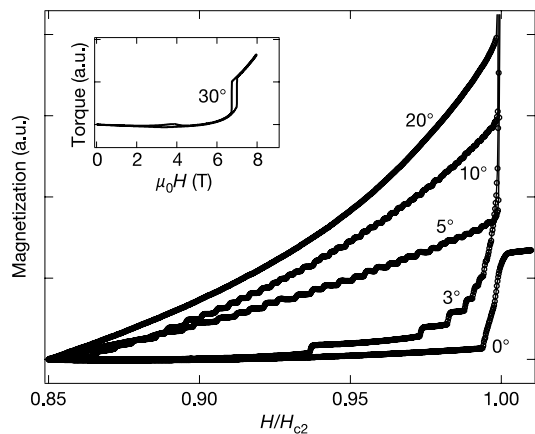


Figure 3 Landau level quantization within the FFLO state. Figure shows magnetization versus field at various field angles with respect to the conducting planes at $T = 25$ mK. The inset shows the cantilever signal for a complete field sweep cycle at 30° with a first order transition visible at 7 T. The measured torque signal, $\tau \sim M \times H$, is a voltage proportional to the gap between the flexible sample plate and a fixed reference plate. The total gap separation is measured as a capacitance using a precision capacitance bridge. The curves in the main panel are collapsed at $0.85H_{c2}$ for clarity. At large angles the magnetization increases smoothly towards H_{c2} , while within $\pm 10^\circ$ of the parallel orientation distinct jumps appear. These jumps represent transitions of the order parameter into higher Landau level states and are a direct consequence of the modulated FFLO phase^{10,11}. We have observed these steps on two different samples using a metal BeCu (13 μm thick) and a metal MP35N (8 μm thick) cantilever which rules out sample and cantilever artefacts. We have also excluded flux avalanches as the source of these jumps. Upon repeated field sweep cycles, the jumps appear at exactly the same values of the applied field and are of uniform step size, whereas flux avalanches show a distribution of jump sizes^{29,30}. No steps in the magnetization were observed at temperatures exceeding T_{FFLO} , confirming their non-systematic origin.

CeCoIn₅ crystallizes in a tetragonal structure with alternating layers of CeIn₃ and CoIn₂. It belongs to the isostructural group of CeMIn₅ ($M = \text{Co, Ir, Rh}$). These materials have attracted considerable attention owing to the interplay of superconductivity and magnetism. CeRhIn₅ has an antiferromagnetic ground state with pressure induced superconductivity above 16 kbar (ref. 17), while CeIrIn₅ and CeCoIn₅ are ambient pressure superconductors. On the basis of these findings, the existence of a quantum critical point has been proposed where the ground state becomes superconducting when antiferromagnetism is suppressed to $T = 0$ K by external or chemical pressure^{18,19}. The proximity to an antiferromagnetic instability also opens up the possibility for spin-fluctuation mediated superconductivity, which has been shown to be more robust in two than three dimensions²⁰. For our experiments, we used crystal platelets with approximate dimensions of $1 \text{ mm} \times 1 \text{ mm} \times 0.1 \text{ mm}$ and masses of the order of 0.25 mg. The single crystals were synthesized from an In flux by combining stoichiometric amounts of Ce and Co. Excess In was removed during preparation with a centrifuge and afterwards by etching in a dilute HCl/HNO₃ aqueous solution¹⁴.

The main panel of Fig. 2 shows measurements of heat capacity as function of applied field, C versus H , for two field orientations with respect to the conducting planes at $T = 250$ mK ($T/T_c = 0.12$). At small angles, when the magnetic field is near the plane parallel orientation ($\theta = 0^\circ$), two phase transitions are present: a lower-field

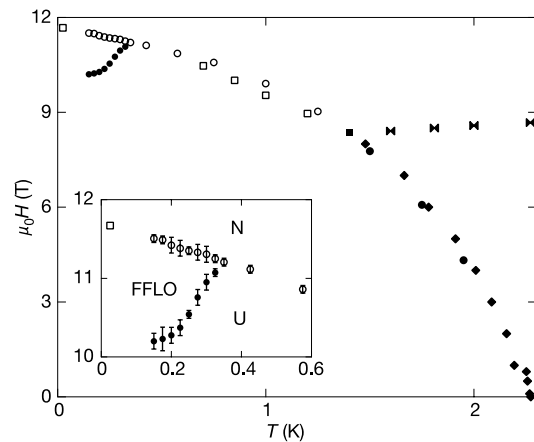


Figure 4 Experimental H - T phase diagram for CeCoIn₅. Transitions were obtained from heat capacity (circles), magnetization (squares) and transport (diamonds) measurements. Open symbols denote first order phase transitions; filled symbols represent second order phase transitions. The external field is applied parallel to the conducting planes. Two distinct phase transitions emerge below $T_{\text{FFLO}} \approx 0.35$ K. The lower, second order phase transition (filled circles) corresponds to the crossover from the uniform superconducting state to the spatially modulated FFLO state, while the upper boundary (open circles) corresponds to a first order phase transition into the normal state. The first order character of the transition extends beyond the tricritical point at T_{FFLO} up to 1.25 K. This specific temperature marks a change in $H_{c2}(T)$ from a first order to a second order phase transition with increasing temperature. The 'bow-tie' symbols indicate a metamagnetic transition that merges with H_{c2} at 1.4 K (ref. 24). It is unknown if the metamagnetic transition has any causal relation to the superconducting-normal phase boundary. The inset shows data below 0.6 K on an expanded scale with the uniform (U) BCS superconducting state and the normal (N) state above $H_{c2}(T)$. The decrease in the lower transition field with decreasing temperature indicates a dominance of negative (antiferromagnetic) exchange interactions according to extensive calculations²⁶. Proximity to an antiferromagnetic quantum critical point should favour this type of interaction. Within the FFLO phase, higher Landau level solutions for the order parameter can occur when both orbital and spin effects play a role. This is seen in the magnetization measurements as a cascade of first order phase transitions at sub-phases with higher orbital momentum (Fig. 1).

second-order phase transition within the superconducting state plus a higher-field first-order phase transition at H_{c2} . We find that this lower-field transition exists only within $\pm 10^\circ$ of the plane parallel field orientation. At larger angles ($\theta = 30^\circ$), there is a single transition at H_{c2} . This transition is still first order, as can be seen from the magnetocaloric data for $\theta = 30^\circ$. A first order transition at H_{c2} for field applied perpendicular to the planes has been by various groups and attributed to a dominance of paramagnetic effects^{21,22}.

The heat capacity measurements are readily explained within the FFLO scenario. The quasi-two-dimensional nature of the crystal inhibits orbital motion of the Cooper pairs in the parallel field orientation. With increasing angle, the orbital component begins to dominate and prevents the non-uniform state from appearing. The higher-order, lower-field transition then marks the uniform superconducting-to-FFLO state boundary. Calculations indeed predict this transition to be of second order²³. Our results provide calorimetric evidence for the FFLO state.

Figure 3 shows torque magnetometry measurements as a function of an applied magnetic field. The main panel comprises 'up' sweeps for five different field orientations between 20° out of plane and plane parallel. Shown are magnetization data versus normalized applied field, H/H_{c2} . The 20° curve displays a smooth increase in magnetization, much as the 30° curve in the inset. A series of steps appears at 10° out of plane, starting above $0.85H_{c2}$ before the superconducting-to-normal transition takes place. As the angle is decreased the steps become more pronounced and their number is reduced, until in the pure paramagnetic limit at 0° orientation these features disappear. It is important to note that identical behaviour is seen for the 'down' sweep in field (not shown for clarity), as well as for all field angles on either side of the field parallel orientation. This cascade signature is seen only within $\pm 10^\circ$ of parallel, the same range of angles for which the lower FFLO phase transition line is observed in the heat capacity measurements. This cascade is a direct consequence of the spatially varying superconducting order parameter promoting higher orbital solutions^{10,11}. As discussed above, higher Landau level states of the order parameter correspond to multiquanta vortex states. This suggests that within the FFLO state, taking into account finite orbital effects, the usual Abrikosov vortices are replaced by multiquanta vortices carrying flux $m\Phi_0$. The character and behaviour of the multiquanta vortices is expected to differ from the single quantum vortices, but no phenomenological description of the field dependence has been put forward at this time.

Figure 4 presents the experimentally determined field–temperature phase diagram for the field orientation parallel to the superconducting planes as derived from magnetization and heat capacity data for CeCoIn₅. We have included resistance measurements from ref. 24 to show consistency with our data and to extend H_{c2} to zero field. For temperatures below approximately 0.35 K, two distinct phase transitions emerge. We identify the temperature $T \approx 0.35$ K as the onset of the FFLO state in CeCoIn₅. The second-order, lower-field transition (Fig. 4, filled circles) marks the crossover from the uniform superconducting state to the spatially varying FFLO state, while the first-order, upper transition (Fig. 4, open symbols) identifies the FFLO-to-normal state transition $H_{c2}(T)$. By using the specific heat jump in zero field of CeCoIn₅ (ref. 15), we estimate a lower bound for the Pauli limiting field to be in the range of 7.6–15.1 T for the plane parallel orientation (see Supplementary Information), significantly higher than expected in the weak coupling, free electron limit. A first order phase transition for temperatures above 0.35 K separates the uniform superconducting state from the normal state. It becomes second order (Fig. 4, filled symbols) for temperatures above approximately 1.25 K. Interestingly, the change in the order of the upper critical transition field roughly coincides with the appearance of a metamagnetic transition (Fig. 4, 'bow-tie' symbols) as seen in ref. 24 for field aligned parallel to the planes, although at this point it has not been determined

what—if any—effect this has on the phase boundary. It is important to note that the most recent magnetotransport measurements on CeCoIn₅ reveal a similar line within the normal state region of the H – T phase diagram for fields applied perpendicular to the planes²⁵. Thus, the metamagnetic transition and the change in the order of H_{c2} shown in Fig. 4 are, unlike the appearance of the FFLO state, independent of field orientation.

Our measurements provide, to our knowledge, the first direct thermodynamic evidence for the existence of the FFLO state. These observations confirm a four-decade-old theory, and open up a new possibility of studying the interaction of magnetism and superconductivity, two states of matter once thought to be mutually exclusive. In this case, however, the formation of magnetic texture in form of the spin aligned Bloch walls effectively lowers the free energy of the system, making it more stable. The realization of the FFLO state in this reduced dimensional system may shed light on magnetically mediated electron pairing and the superconducting mechanism of other layered systems, such as high-temperature superconductors and molecular superconductors. □

Received 30 December 2002; accepted 12 June 2003; doi:10.1038/nature01842.

- Onnes, H. K. The resistance of pure mercury at helium temperatures. *Comm. Phys. Lab. Univ. Leiden*. No. B 120, 3 (1911).
- Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of superconductivity. *Phys. Rev.* **108**, 1175–1204 (1957).
- Werthamer, N. R., Helfand, E. & Hohenberg, P. C. Temperature and purity dependence of the superconducting critical field, H_{c2} . III. Electron spin and spin-orbit effects. *Phys. Rev.* **147**, 295–302 (1966).
- Clogston, A. M. Upper limit for critical field in hard superconductors. *Phys. Rev. Lett.* **9**, 266–267 (1962).
- Maki, K. Effect of Pauli paramagnetism on magnetic properties of high-field superconductors. *Phys. Rev.* **148**, 362–369 (1966).
- Fulde, P. & Ferrell, R. A. Superconductivity in a strong spin-exchange field. *Phys. Rev.* **135**, A550–A563 (1964).
- Larkin, A. I. & Ovchinnikov, Y. N. Inhomogeneous state of superconductors. *Sov. Phys. JETP* **20**, 762–769 (1965).
- Norman, M. R. Existence of the FFLO state in superconducting UPd₂Al₃. *Phys. Rev. Lett.* **71**, 3391 (1993).
- Agterberg, D. F. & Yang, K. The effect of impurities on Fulde-Ferrell-Larkin-Ovchinnikov superconductors. *J. Phys. Condens. Matter* **13**, 9259–9270 (2001).
- Buzdin, A. I. & Brison, J. P. New solutions for the superconducting order parameter in a high magnetic field. *Phys. Lett. A* **218**, 359–366 (1996).
- Buzdin, A. I. & Brison, J. P. Non-uniform state in 2D superconductors. *Europhys. Lett.* **35**, 707–712 (1996).
- Shimahara, H. & Rainer, D. Crossover from vortex states to the Fulde-Ferrell-Larkin-Ovchinnikov state in two-dimensional s- and d-wave superconductors. *J. Phys. Soc. Jpn* **66**, 3591–3599 (1997).
- Houzet, M. & Buzdin, A. Structure of the vortex lattice in the Fulde-Ferrell-Larkin-Ovchinnikov state. *Phys. Rev. B* **63**, 184521 (2001).
- Petrovic, C. *et al.* Heavy-fermion superconductivity in CeCoIn₅ at 2.3 K. *J. Phys. Condens. Matter* **13**, L337–L342 (2001).
- Ikeda, S. *et al.* Unconventional superconductivity in CeCoIn₅ studied by the specific heat and magnetization measurements. *J. Phys. Soc. Jpn* **70**, 2248–2251 (2001).
- Movshovich, R. *et al.* Unconventional superconductivity in CeIrIn₅ and CeCoIn₅: Specific heat and thermal conductivity studies. *Phys. Rev. Lett.* **86**, 5152–5155 (2001).
- Hegger, H. *et al.* Pressure-induced superconductivity in quasi-2D CeRhIn₅. *Phys. Rev. Lett.* **84**, 4986–4989 (2000).
- Sachdev, S. Quantum criticality: Competing ground states in low dimensions. *Science* **288**, 475–480 (2000).
- Nicklas, M. *et al.* Response of the heavy-fermion superconductor CeCoIn₅ to pressure: Roles of dimensionality and proximity to a quantum-critical point. *J. Phys. Condens. Matter* **13**, L905–L912 (2001).
- Monthoux, P. & Lonzarich, G. G. Magnetically mediated superconductivity in quasi-two and three dimensions. *Phys. Rev. B* **63**, 054529 (2001).
- Tayama, T. *et al.* Unconventional heavy-fermion superconductor CeCoIn₅: dc magnetization study at temperatures down to 50 mK. *Phys. Rev. B* **65**, 180504 (2002).
- Bianchi, A. *et al.* First-order superconducting phase transition in CeCoIn₅. *Phys. Rev. Lett.* **89**, 137002 (2002).
- Matsuo, S., Higashitani, S., Nagato, Y. & Nagai, K. Phase diagram of the Fulde-Ferrell-Larkin-Ovchinnikov state in a three-dimensional superconductor. *J. Phys. Soc. Jpn* **67**, 280–289 (1998).
- Murphy, T. P. *et al.* Anomalous superconductivity and field-induced magnetism in CeCoIn₅. *Phys. Rev. B* **65**, 100514 (2002).
- Paglione, J. *et al.* Field-induced Fermi liquid state in CeCoIn₅. *Phys. Rev. Lett.* (submitted); preprint at (<http://xxx.lanl.gov/cond-mat/0212502>) (2002).
- Burkhardt, H. & Rainer, D. Fulde-Ferrell-Larkin-Ovchinnikov state in layered superconductors. *Ann. Phys.* **3**, 181–194 (1994).
- Hannahs, S. T. & Fortune, N. A. Heat capacity cell for angular measurements in high magnetic fields. *Physica B* **329–333**, 1586–1587 (2003).
- Fortune, N. A. *et al.* High magnetic field corrections to resistance thermometers for low temperature calorimetry. *Rev. Sci. Instrum.* **71**, 3825–3830 (2000).

29. Zieve, R. J. *et al.* Vortex avalanches at one thousandth the superconducting transition temperature. *Phys. Rev. B* **53**, 11849–11854 (1996).
30. Mola, M. M., Hill, S., Qualls, J. S. & Brooks, J. S. Magneto-thermal instabilities in an organic superconductor. *Int. J. Mod. Phys. B* **15**, 3353–3356 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. L. Sarrao for providing the samples, C. C. Agosta for discussions, V. Williams, D. McIntosh, J. Farrell and J. Kosakowski for technical assistance, and N. Malkovich for the Parmax(R) 1200 used in our heat capacity cell. This work was supported by the National Science Foundation, the State of Florida, and NHMFL Visiting Scientist and In-House Research Programs.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.A.R. (radovan@magnet.fsu.edu).

Geochemical evidence for efficient aquifer isolation over geological timeframes

Bernard Marty^{1,2}, Sarah Dewonck^{1*} & Christian France-Lanord¹

¹Centre de Recherches Pétrographiques et Géochimiques (CRPG), 15 Rue Notre-Dame des Pauvres, and ²Ecole Nationale Supérieure de Géologie, Rue du Doyen Marcel Roubault, BP 40, 54501 Vandœuvre lès Nancy Cedex, France

* Present address: ANDRA, BP 9, 55290 Bure, France

Aquitards—layers of rock having low permeability—have been suggested as potential long-term reservoirs for toxic materials such as nuclear or chemical waste. But information about the isolation properties of aquitard layers is essential to evaluate whether they can indeed be used safely as reservoirs. Here we investigate the long-term mobility of groundwaters between two aquifers surrounding an aquitard layer in the eastern recharge area of the Paris basin, France, using helium isotopes as a geochemical tracer. The deeper Trias sandstone aquifer, which lies above the crystalline basement, accumulates radiogenic ⁴He and primordial ³He from large regions of the crust and mantle at rates comparable to the degassing of the whole crust¹ and of mid-ocean ridges². We show that the overlying carbonate Dogger aquifer, which is separated from the Trias aquifer by an aquitard layer consisting of a ~600 m succession of shales and clays, is stagnant and has been extremely well isolated from the Trias over the past several million years. This finding, together with previous studies at the centre of the Paris basin^{3,4}, shows that diffusive mass transfer across aquitards is negligible and that cross-formational flow in basins takes place preferentially in faulted areas.

The long-term (10⁵–10⁷ yr) impact of human activity is a major environmental issue, in particular regarding confinement of toxic material in geological formations. Certain shale- and clay-rich successions which often act as aquitards in sedimentary basins are under consideration as potential host formations for underground storage of nuclear waste. Some authors have argued that aquifers are isolated from other formations (see ref. 5), whereas others have proposed that aquifer waters are open systems¹ (see ref. 6 for discussion). The mass transfer properties of aquitards over time-scales of several millions of years (Myr) are not well documented and rely mostly on permeation and diffusion measurements at laboratory-accessible space and timescales or on modelling

approaches; the former give contrasting results^{7,8} while the latter are limited in their capacity to represent the effects of natural media heterogeneities over such long durations. Evaluating the mobility of deep groundwaters requires the use of natural geochemical tracers, among which noble gases are particularly useful because of their chemical inertness and multiple origins (air, mantle, crust)⁶. We present here the results of a noble gas study of two superposed aquifers in the eastern Paris basin, France, that support efficient aquifer isolation by the intermediate aquitard over several Myr.

The Paris basin provides an excellent framework for testing geochemical tracers. It is a well-documented multi-layered basin that developed by thinning of the underlying crust. The major recharge area, the Vosges (Lorraine region, Fig. 1) and the Morvan, have been uplifted by Alpine deformation and have probably acted as recharge areas during at least the second half of the Tertiary period⁹. The general flow direction of groundwaters inferred from hydrology and water geochemistry is from the south and southeast to the north and northwest^{3,4,10}. Among the different aquifers present in the basin (Fig. 1), the Dogger (oolithic limestone of marine origin intercalated with marls) and the Trias (sandstone) aquifers have been extensively drilled for drinking water, geothermal energy and oil recovery purposes. Hydrological modelling¹⁰ and stable isotope³ and noble gas data¹¹ studies of fluids from the central part of the Dogger and Trias aquifers showed the occurrence of old groundwaters (~10⁶ yr; refs 3, 12, 13) that were recharged during a period warmer than the present one. The present study focuses on the eastern border where the Dogger and the Trias groundwaters are separated by about 600 m of upper Trias and Lias shales and clays and the Dogger is covered by ~100 m of Callovo–Oxfordian shales and clays. Eighteen drill holes tapping Trias groundwaters and five drill holes tapping the Dogger aquifer (of which MSE101 was drilled for scientific purposes by the French Nuclear Waste Agency, ANDRA) were sampled and analysed for stable isotopes, ¹⁴C and noble gases (Fig. 1, Table 1).

Helium contents and isotope ratios together with ¹⁴C activities indicate that the Trias aquifer receives fluids from the underlying basement. ¹⁴C dating of Trias groundwaters indicate that Trias waters were recharged ≤30,000 yr ago (Table 1) and are flowing at an average velocity of 3.7 ± 1.3 m yr⁻¹ (Fig. 2). Stable isotope data are in agreement with this, because the δD and δ¹⁸O values lower than present-day precipitation values (Table 1) indicate recharge during a colder period¹⁴. The concentrations of dissolved helium are largely in excess of those expected for recharge water equilibrium with the atmosphere (Table 1). These excesses cannot be due to contribution of radiogenic helium, ⁴He, produced in the rock porosity and injected into the formation water (in situ production) because the amount of ⁴He generated during the water residence time¹⁵ (computed with a ⁴He accumulation rate in water of 2 × 10⁻¹⁶ mol g⁻¹ yr⁻¹; ref. 11) could only account for ≤1% of observed ⁴He. Explanation of the high He contents of Trias groundwaters requires contribution of helium from a source outside the aquifer. An exotic He input is independently indicated by groundwater ³He/⁴He ratios (mean value = 0.4 ± 0.1 R_a where R_a is the atmospheric value) which are higher than those expected for local radiogenic production in sediments or crust^{4,11}. The ³He/⁴He ratios indicate instead a contribution of mantle-derived helium (³He/⁴He = 8 ± 1 R_a for the upper mantle¹⁶) in Lorraine.

The ³He and ⁴He groundwater contents increase with distance from the recharge area (Fig. 2), allowing computation of the He isotope fluxes entering the aquifer based on He isotope accumulation rates along a flow line, Trias aquifer thickness (300 m on average) and rock porosity (15%)^{11,17}. The computed mean ⁴He flux of 4.9 ± 2.1 × 10⁻⁶ mol m⁻² yr⁻¹ is comparable with, although higher than, the crustal flux of 1.5 × 10⁻⁶ mol m⁻² yr⁻¹ corresponding to the radiogenic production in the entire continental crust^{1,16}. For comparison, the ⁴He crustal flux in the tectonically inactive centre of the Paris basin is lower by one order of magnitude

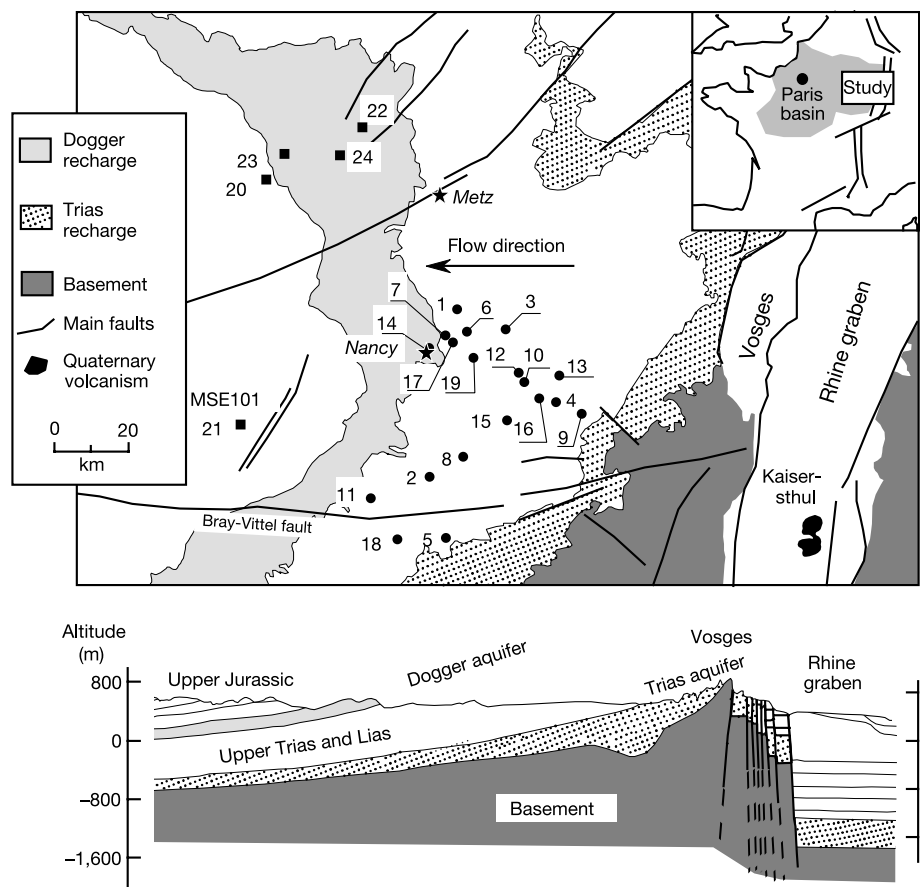


Figure 1 Schematic map of the eastern area of the Paris basin, France. Filled circles and filled squares represent the location of the sampled wells tapping the Trias aquifer and the Dogger aquifer respectively (numbers refer to site names given in Table 1).

Table 1 Stable isotope ^{14}C and noble gas data for Lorraine groundwaters

Sampling site	Well number	Distance recharge (km)	Depth (m below surface)	$\delta^{18}\text{O}$ (‰ versus SMOW)	δD (‰ versus SMOW)	A^{14}C (pmc)	$t_{14\text{C}}$ (yr BP)	^4He ($10^{-9}\text{ mol g}^{-1}$)	$^3\text{He}/^4\text{He}$ (R_a)	$^4\text{He}/^{20}\text{Ne}$
Trias										
Agincourt	1	90	834	-9.4	-68.6	1.7 ± 0.1	24,048	2.31	0.41 ± 0.01	198
Baudricourt	2	40	460	-9.9	-67.4	7.4 ± 0.2	16,597	1.16	0.32 ± 0.01	74
Bénamont	3	85	549	-9.2	-66.0	2.3 ± 0.2	22,505	0.89	0.37 ± 0.01	87
Chenevières	4	50	453	-9.7	-70.4	-	-	1.49	0.27 ± 0.01	89
Escles	5	15	80	-8.5	-58.0	77.0 ± 0.5	2,284	0.01	0.32 ± 0.06	1.1
Essey	7	85	681	-9.6	-71.2	1.7 ± 0.2	24,392	3.27	0.37 ± 0.01	257
Florémont	8	45	-	-9.4	-65.6	4.6 ± 0.2	15,996	1.14	-	70
Gélacourt	9	45	225	-10.0	-71.4	-	-	0.28	0.25 ± 0.01	26
Hériménil	10	65	412	-9.6	-72.0	5.0 ± 0.1	14,417	1.84	0.41 ± 0.01	169
La Neuveville	11	35	279	-9.7	-66.3	1.8 ± 0.1	27,950	1.62	0.30 ± 0.01	100
Lunéville	12	70	451	-9.6	-69.8	3.2 ± 0.2	21,002	1.52	0.64 ± 0.01	138
Manonviller	13	50	410	-10.0	-71.1	-	-	0.13	0.35 ± 0.02	11
Nancy Thermal	14	80	850	-9.3	-67.1	-	-	28.5	0.29 ± 0.01	2,180
Rozelleures	15	55	500	-9.6	-64.1	5.8 ± 0.1	13,190	2.32	0.40 ± 0.02	155
St Clément	16	65	319	-9.5	-67.9	7.3 ± 0.2	12,981	2.00	0.30 ± 0.02	185
Tomblaine	17	80	708	-9.4	-69.0	1.5 ± 0.2	24,370	4.57	0.37 ± 0.01	340
Valfroicourt	18	25	150	-8.6	-59.2	38.7 ± 0.4	6,052	0.57	0.28 ± 0.01	60
Varangéville	19	75	590	-9.5	-69.0	1.8 ± 0.1	23,765	2.30	0.38 ± 0.02	191
Dogger										
Etain	20	-	288	-8.5	-58.3	9.2 ± 0.1	19,700	0.024	0.38 ± 0.01	2.0
MSE101	21	35	650-675	-6.9	-46.6	2.5 ± 0.1	30,500	7.27	0.023 ± 0.002	599
Piennes	22	-	240	-8.2	-56.7	7.0 ± 0.1	21,900	0.016	0.62 ± 0.02	0.68
Rouvres	23	-	105	-8.7	-59.6	2.5 ± 0.1	30,400	0.39	0.040 ± 0.002	32
Valleroy	24	-	165	-7.7	-51.2	35.7 ± 0.1	8,520	0.0055	0.73 ± 0.02	0.39
Surface water				-7.7 to -8.3	-50 to -60			0.0021	1.1	0.254

^{14}C ages of Trias and Dogger waters are computed following method from ref. 23 and the conventional method, respectively. The surface water composition is an average of the Moselle river in Lorraine for stable isotopes, and is air-saturated water (ASW) at 10°C for helium. The occurrence of ^{14}C in MSE101 water is likely to result from residual contamination by drilling fluids not washed out completely^{29,30}. A^{14}C is the measured activity of ^{14}C . pmc, per cent modern carbon.

($\sim 2 \times 10^{-7} \text{ mol m}^{-2} \text{ yr}^{-1}$; ref. 12). The eastern border of the basin is limited by the Rhine graben (Fig. 1), which presents an enhanced heat flow up to 200 mW m^{-2} compared to $\leq 100 \text{ mW m}^{-2}$ in the basin centre¹⁸ and under which asthenospheric upwelling (Moho depth at 24 km) has been detected¹⁹. Hence the ^4He flux at the active eastern margin of the Paris basin, which is a factor of three higher than the steady-state crustal ^4He flux, may represent the purging out of radiogenic He which accumulated in a quiescent crust between crustal deformation and fracturation. This comparison strongly suggests that crustal degassing is a discontinuous, and not a steady-state¹, process, which is primarily driven by changes in the tectonic regime.

The mantle ^3He flux entering the Trias aquifer is $(2.5 \pm 1.1) \times 10^{-12} \text{ mol m}^{-2} \text{ yr}^{-1}$, two orders of magnitude higher than the bulk continental crust flux ($\sim 4 \times 10^{-14} \text{ mol m}^{-2} \text{ yr}^{-1}$; ref. 20), and comparable to the global ^3He flux from the mantle of $2 \times 10^{-12} \text{ mol m}^{-2} \text{ yr}^{-1}$; refs 2, 20. This tectonic situation is analogous to extension and magmatism taking place at mid-ocean ridges (MOR), the main source of mantle ^3He . Magmatism under the Rhine graben is indicated by $^3\text{He}/^4\text{He}$ ratios up to $1.7 R_a$ in mineral springs from Alsace²¹, and the most recent volcanic activity in the Rhine graben occurred at the Kaiserstuhl volcano 16 Myr ago (Fig. 1). This study allows us to estimate a mantle-derived magma generation rate that can account for the observed ^3He flux. The

maximum distance from the graben axis of Trias wells showing excesses of ^3He flux is 120 km (Fig. 1). Going towards the centre of the basin, the nearest Trias wells, which are sampling oil reservoirs and are located in Champagne 340 km from the graben axis, present $^3\text{He}/^4\text{He}$ ratios of $0.02 R_a$ (ref. 4), indicating limited, if any, mantle ^3He contribution. Assuming that the lateral extent of the ^3He mantle flux is ~ 200 km on both sides of the Rhine graben, the ^3He flux of a 100-km-long section of the Rhine graben is $\sim 0.1 \text{ mol yr}^{-1}$. This is one order of magnitude lower than the mean ^3He flux of a 100-km-long MOR section ($\sim 2.5 \text{ mol yr}^{-1}$, computed from the ^3He MOR flux of $\sim 1,000 \text{ mol yr}^{-1}$ (ref. 2) and a total MOR length of 40,000 km), consistent with a significantly cooler thermal regime of continental rifting compared to the oceanic one. Assuming a ^3He asthenospheric mantle content of $1.5 \times 10^{-15} \text{ mol g}^{-1}$ (ref. 22), the magma generation rate for the 100-km-long Rhine graben section needed to account for its ^3He flux is as low as $\sim 10^{-3} \text{ km}^3 \text{ yr}^{-1}$. For comparison, the last eruption in the Rhine graben north of the study area, that of Laacher See in the Eifel region (Germany) 11,000 yr ago, required the generation and differentiation of a non-erupted basic magma volume of 50 km^3 (ref. 23).

In contrast, the $^3\text{He}/^4\text{He}$ isotopic ratios of the least atmosphere-contaminated waters from the Dogger aquifer ($0.02 R_a$ for MSE101 and $0.04 R_a$ for Rouvres-en-Woëvre, Table 1) are low and similar to values expected for radiogenic production in the Dogger and underlying Lias sediments ($0.02\text{--}0.03 R_a$; ref. 11). Furthermore, the ^3He contents of the Dogger waters ($(1.3\text{--}23) \times 10^{-17} \text{ mol g}^{-1}$) are 1–3 orders of magnitude lower than those observed in the westernmost waters of the Trias ($(1.3\text{--}11) \times 10^{-15} \text{ mol g}^{-1}$). Hence the signal of mantle-derived helium-3 identified in the Trias waters has not reached the Dogger waters, demonstrating the isolating property of the upper Trias–Lias aquitard. As the eastern Dogger waters are efficiently isolated from He isotope contributions from the Trias and from the basement, appropriate source(s) of He must be found for this aquifer, the possibilities being the adjacent aquitards, the upper Trias–Lias at the bottom and the Callovo–Oxfordian at the top. For the Etain and Rouvres-en-Woëvre Dogger sites (Fig. 1), ^{14}C ages together with ^4He contents (Table 1) permit calculation of ^4He accumulation rates of $0.4 \times 10^{-15} \text{ mol g}^{-1} \text{ yr}^{-1}$ and $12 \times 10^{-15} \text{ mol g}^{-1} \text{ yr}^{-1}$ respectively, which are roughly consistent with the ^4He accumulation rate of $\sim 1 \times 10^{-15} \text{ mol g}^{-1} \text{ yr}^{-1}$ corresponding to radioactive contribution from the decays of U and Th present in the Dogger and aquitard rocks¹¹. The residence time of MSE101 water computed using this latter value is ~ 7 Myr, which probably represents a lower limit because the isolating property of shales and clays deduced from ^3He also prohibits efficient transfer of ^4He from adjacent aquitards. A long residence time for the Dogger water is fully consistent with the stable isotope ratios which show enrichments in D and ^{18}O relative to modern waters, indicating warmer environmental conditions and therefore a pre-Quaternary recharge. Such a long residence time implies that the Dogger water is stagnant, possibly since several Myr ago, in sharp contrast with the underlying Trias water which is flowing, possibly since the Miocene epoch, at a present-day velocity of 4 m yr^{-1} .

This study demonstrates that mass transfer across aquitards can be extremely limited on geological timeframes, which is consistent with available experimental data. The minimum time interval for a He atom to diffuse through the upper Trias and Lias is h^2/D_a , where h is the thickness of the impermeable layer (600 m) and D_a is the effective diffusion coefficient integrated over the thickness of the aquitard. A D_a value of $6 \times 10^{-11} \text{ m}^2 \text{ s}^{-1}$ has been recently determined for a Callovo–Oxfordian aquitard sample⁸, resulting, if representative of the upper Trias–Lias aquitard, in an extraordinarily long (190 Myr) diffusive transfer. However, chemical and isotopic data from rocks and formation waters sampled on a basin scale show the occurrence of cross-formational flow from the Trias aquifer to the Dogger aquifer in the centre of the Paris basin^{3,4,24} and that this

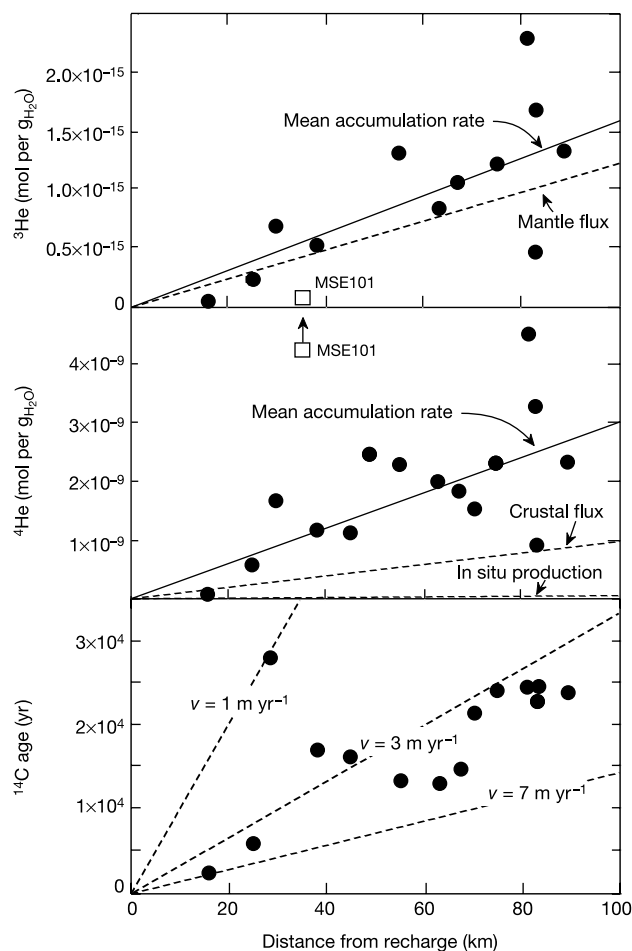


Figure 2 Evolution of ^{14}C ages, ^4He and ^3He contents for Trias groundwaters (black dots) as a function of approximate distance from the recharge area. Open squares represent Dogger groundwaters sampled in the scientific drillhole MSE101 (see text for significance of evolution lines).

flow must have been convective rather than diffusive in order to preserve the characteristics of water input from the Trias. According to the distribution of salinity and other geochemical tracers in the Dogger aquifer, cross-formation flow probably took place via the major Hercynian fault and fracture systems that run through the centre of the Paris basin^{4,24}. We conclude that convection through geological accidents is probably the main mode of cross-formation mass transfer in basins because diffusive transfer appears to be extremely limited on a basin scale. The detailed role of diffusion on a local scale will need to be assessed in the future by experiments aimed at documenting the migration of highly mobile elements in aquitards. □

Methods

Stable isotopes analysis was performed at the CRPG following reduction over uranium at 800 °C for hydrogen and isotopic exchange between water and CO₂ for oxygen. As the dissolved species content is low, no correction for the salinity was required. Uncertainties are about ±2‰ and ±0.2‰ for the D/H and ¹⁸O/¹⁶O ratios, expressed as δ values relative to V-SMOW. ¹⁴C activities were measured by Accelerator Mass Spectrometry at the University of Arizona at Tucson. ¹⁴C activities have been corrected for secular variation of ¹⁴C activity²⁵ and for dilution by “dead” carbon using the procedure of Ingerson and Pearson²⁶, using δ¹³C values also measured for these samples. A more rigorous correction would need knowledge of carbon speciation in groundwaters²⁷, which is not available for this study. However, for a Palaeozoic sandstone aquifer in Western Flanders (Belgium and France) for which the context is comparable to the Vosges sandstone aquifer of this study, the isotopic correction gave water ages comparable to those obtained using the chemical model²⁷. In the case of Dogger samples, as the Ingerson and Pearson correction does not work for carbonated aquifers, we instead use conventional ¹⁴C ages. For noble-gas analysis, copper tubes containing the water sampled at wellheads under pressure were connected to a high-vacuum stainless steel extraction-purification line in CRPG, Nancy. Helium and neon were then separated cryogenically and analysed sequentially by static mass spectrometry²⁸.

Received 27 February; accepted 31 July 2003; doi:10.1038/nature01966.

1. Torgersen, T. & Clarke, W. B. Helium accumulation in groundwater. I: an evaluation of sources and the continental flux of crustal ⁴He in the Great Artesian Basin, Australia. *Geochim. Cosmochim. Acta* **49**, 1211–1218 (1985).
2. Craig, H., Clarke, W. B. & Beg, M. A. Excess ³He in deep waters on the East Pacific Rise. *Earth Planet. Sci. Lett.* **26**, 125–132 (1975).
3. Matray, J. M., Lambert, M. & Fontes, J. C. Stable isotope conservation and origin of saline waters from the Dogger aquifer of the Paris basin, France. *Appl. Geochem.* **9**, 297–309 (1994).
4. Pinti, D. L. & Marty, B. Noble gases in crude oils from the Paris basin, France—Implications for the origin of fluids and constraints on oil–water interactions. *Geochim. Cosmochim. Acta* **59**, 3389–3404 (1995).
5. Tolstikhin, I. N., Lehmann, B. E., Loosli, H. H. & Gautschi, A. Helium and argon isotopes in rocks, minerals, and related groundwaters: A case study in northern Switzerland. *Geochim. Cosmochim. Acta* **60**, 1497–1514 (1996).
6. Ballentine, C. J., Burgess, R. & Marty, B. Tracing fluid origin, transport and interaction in the crust. *Rev. Mineral.* **47**, 539–614 (2002).
7. Ohsumi, T. & Horibe, Y. Diffusivity of He and Ar in deep-sea sediments. *Earth Planet. Sci. Lett.* **70**, 61–68 (1984).
8. Rebou, V., Billiotte, J., Deveughele, M., Jambon, A. & le Guen, C. Molecular diffusion in water-saturated rocks: a new experimental method. *J. Contam. Hydrol.* **28**, 71–93 (1997).
9. Mégnien, C. Synthèse géologique du Bassin de Paris. *Mém. BRGM* **101**, 1–466 (1980).
10. Menjoz, A., Lambert, M. & Matray, J. M. Flow of formation waters in the Jurassic of the Paris basin, France. *Phil. Trans. R. Soc. Lond. A* **344**, 141–156 (1994).
11. Pinti, D. L. & Marty, B. in *Dating and Duration of Fluid Flow and Fluid–Rock Interaction* (ed. Parnell, J.) 53–68 (Geol. Soc. Spec. Publ., London, 1998).
12. Marty, B., Torgersen, T., Meynier, V., O’Nions, R. K. & de Marsily, G. Helium isotope fluxes and groundwater ages in the Dogger Aquifer, Paris basin. *Wat. Resour. Res.* **29**, 1025–1035 (1993).
13. Castro, M. C., Jambon, A., de Marsily, G. & Schlosser, P. Noble gases as natural tracers of water circulation in the Paris basin. 1. Measurements and discussion of their origin and mechanisms of vertical transport in the basin. *Wat. Resour. Res.* **34**, 2443–2466 (1998).
14. Blavoux, B. & Olive, P. Radiocarbon dating of groundwater of the aquifer confined in the lower Triassic sandstones of the Lorraine region, France. *J. Hydrol.* **54**, 167–183 (1981).
15. Marine, I. W. The use of naturally occurring He to estimate groundwater velocities for studies of geological storage waste. *Wat. Resour. Res.* **15**, 1130–1136 (1979).
16. Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* (Cambridge Univ. Press, Cambridge, 2002).
17. Andrews, J. N. et al. A radiochemical, hydrochemical and dissolved gas study of groundwaters in the Molasse basin of Upper Austria. *Earth Planet. Sci. Lett.* **73**, 317–332 (1985).
18. Ungemach, T. Les ressources géothermiques de la Communauté Européenne. Etat de l’exploration, perspectives de développement. *Hydrologie-Géologie de l’Ingénieur* **1**, 11–42 (1984).
19. Rousset, D., Bayer, R., Guillon, D. & Edel, J.-B. Structure of the southern Rhine Graben from gravity and reflection seismic data (ECORS-DEKORP program). *Tectonophysics* **221**, 135–153 (1993).
20. Torgersen, T. Terrestrial helium degassing fluxes and the atmospheric helium budget: Implications with respect to the degassing processes of continental crust. *Chem. Geol.* **79**, 1–14 (1989).
21. Griesshaber, E., O’Nions, R. K. & Oxburgh, E. R. Helium and carbon isotope systematics in crustal fluids from the Eifel, the Rhine graben and Black Forest, F.R.G. *Chem. Geol.* **99**, 213–235 (1992).
22. Kellogg, L. H. & Wasserburg, G. J. The role of plumes in mantle helium fluxes. *Earth Planet. Sci. Lett.* **99**, 276–289 (1990).

23. Wörner, G. & Schmincke, H. U. Petrogenesis of the zoned Laacher See Tephra. *J. Petrol.* **25**, 836–851 (1984).
24. Worden, R. H. & Matray, J. M. Cross formational flow in the Paris basin. *Basin Res.* **7**, 53–66 (1995).
25. Bard, E., Hamelin, B., Fairbanks, R. G. & Zindler, A. Calibration of the ¹⁴C timescale over the past 30,000 years using mass spectrometric U–Th ages from Barbados corals. *Nature* **345**, 405–409 (1990).
26. Ingerson, E. & Pearson, F. J. in *Recent Researches in the Fields of Hydrology, Atmosphere and Nuclear Geochemistry* (eds Miyake, Y. & Koyama, T.) 263–283 (Maruzen, Tokyo, 1964).
27. Fontes, J. C. & Garnier, J. M. Determination of the initial ¹⁴C activity of total dissolved carbon: A review of existing models and a new approach. *Wat. Resour. Res.* **15**, 399–413 (1979).
28. Richard, D., Marty, B., Chaussidon, M. & Arndt, N. Helium isotopic evidence for a lower mantle component in depleted Archean Komatiite. *Science* **273**, 93–95 (1996).
29. Report B RP 0ANT 01–010/B, 1–158 (ANDRA, Fontenay-Aux-Roses, 2002).
30. Kloppman, W., Matray, J. M. & Aranyosy, J. F. Contamination of deep formation waters by drillings fluids: correction of the chemical and isotopic composition and evaluation of errors. *Appl. Geochem.* **16**, 1083–1096 (2001).

Acknowledgements This study was supported by the French Nuclear Waste Agency (ANDRA) and by the Région Lorraine. We are grateful to all institutions and individuals who made sampling possible. Analytical support was provided by G. Sauder, P. Robert and L. Zimmermann. We appreciated comments from D. Hilton.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.M. (bmarty@crpg.cnrs-nancy.fr).

Laboratory models of the thermal evolution of the mantle during rollback subduction

C. Kincaid¹ & R. W. Griffiths²

¹Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island 02882, USA

²Research School of Earth Sciences, Australian National University, Canberra, 0200 Australian Capital Territory

The subduction of oceanic lithosphere plays a key role in plate tectonics, the thermal evolution of the mantle and recycling processes between Earth’s interior and surface. Information on mantle flow, thermal conditions and chemical transport in subduction zones come from the geochemistry of arc volcanoes^{1–3}, seismic images^{4,5} and geodynamic models^{6–10}. The majority of this work considers subduction as a two-dimensional process, assuming limited variability in the direction parallel to the trench. In contrast, observationally based models increasingly appeal to three-dimensional flow associated with trench migration and the sinking of oceanic plates with a translational component of motion¹¹ (rollback). Here we report results from laboratory experiments that reveal fundamental differences in three-dimensional mantle circulation and temperature structure in response to subduction with and without a rollback component. Without rollback motion, flow in the mantle wedge is sluggish, there is no mass flux around the plate and plate edges heat up faster than plate centres. In contrast, during rollback subduction flow is driven around and beneath the sinking plate, velocities increase within the mantle wedge and are focused towards the centre of the plate, and the surface of the plate heats more along the centreline.

Schematic models of three-dimensional (3D) circulation in subduction zones have been developed from seismic data^{12,13}, including depictions of vigorous flow around plate edges¹⁴. Such

flows induced by rollback subduction may also provide a mechanism for transporting geochemically distinct mantle from the ocean side of the slab into the wedge^{15–17} (Fig. 1) and producing anomalous melting patterns at plate edges^{18,19}. Mass conservation requires that mantle displaced from the ocean side of the plate (Fig. 1a) moves towards the wedge either around²⁰ or beneath the sinking plate. When limited to a two-dimensional (2D) vertical cross-section, the return flow caused by rollback can occur only beneath the plate^{9,10,21}, steepening the trajectories of return flow into the arc-wedge corner^{9,10,22}. Laboratory models of 3D rollback subduction have documented different styles of slab sinking and deformation at 670 km (refs 23–25) and demonstrated the importance of return flow^{23,26}. However, no previous models have considered the 3D response of the mantle in terms of both temperatures and flow patterns.

In our experiments, the upper mantle is simulated with glucose syrup and the subducting plate is represented with a phenolic-resin sheet that is forced to sink into the glucose along prescribed trajectories (Fig. 1b). The range in sinking styles seen in previous experiments with dynamic slabs^{23–25} is produced using pistons which control downdip (U_D) and translational (U_T) plate motion, and steepen the plate's dip angle with time (θ_t) (Fig. 1b; see Methods). The important parameters are the relative magnitudes

of each of these plate motions, and the initial thickness of the thermal boundary layer (TBL) beneath the overriding plate (Fig. 1b; see Methods).

Length and time scales in the laboratory models are related to the mantle through the Peclet number:

$$Pe = U_D D / \kappa \quad (1)$$

The thermal diffusivity, κ , of the laboratory fluid and phenolic-resin plate are both $\sim 0.001 \text{ cm}^2 \text{ s}^{-1}$. The corresponding mantle value is $0.01 \text{ cm}^2 \text{ s}^{-1}$. A length scale, D , is defined as the thickness of the laboratory plate (2.5 cm) corresponding to 80-km-thick lithosphere. By equating Pe for the laboratory and the mantle, plate rates of $2\text{--}10 \text{ cm min}^{-1}$ in the laboratory correspond to mantle values of $4\text{--}20 \text{ cm yr}^{-1}$. Temperatures are scaled by assuming the initial 15°C plate–fluid difference corresponds to the estimated temperature difference ($1,050^\circ \text{C}$)⁹ between the ambient mantle and the surface of the oceanic plate at a depth of 20 km, approximately the elastic thickness of the overriding plate. Thus a 1°C difference in the laboratory corresponds to 70°C in the mantle.

We select three cases to demonstrate basic differences in downdip versus rollback subduction from a larger set of experiments covering a range in plate sinking rates and styles (Table 1). Flow patterns produced within the fluid for a case of purely downdip sinking ($U_D = 2 \text{ cm min}^{-1}$) versus with rollback ($U_D = U_T = 2 \text{ cm min}^{-1}$) are shown in Fig. 2. For the downdip case (Fig. 2a) fluid motion in a near-surface horizontal plane is uniformly inward, or towards the surfaces of the plate. Velocities are $0.3\text{--}0.35 U_D$ and do not vary significantly in the y direction. Close to the edge ($y = W/2$), fluid trajectories bend around to orientations that are parallel to the overall strike of the trench. Streak patterns in a vertical plane ($x\text{--}z$; $y = 0$) show that flow in the wedge is towards the plate's upper surface at shallow depths before turning downward with the plate. Velocities are $\sim 0.35 U_D$ at shallow depths ($< 3 \text{ cm}$) and fall to $0.2\text{--}0.3 U_D$ for depths of $3\text{--}9 \text{ cm}$. The circulation cell within the wedge is a transient feature related to the initiation of subduction. Motion in the wedge evolves to a typical pattern of 2D corner flow with no mass flux between the ocean- and wedge-side of the fluid^{9,21}.

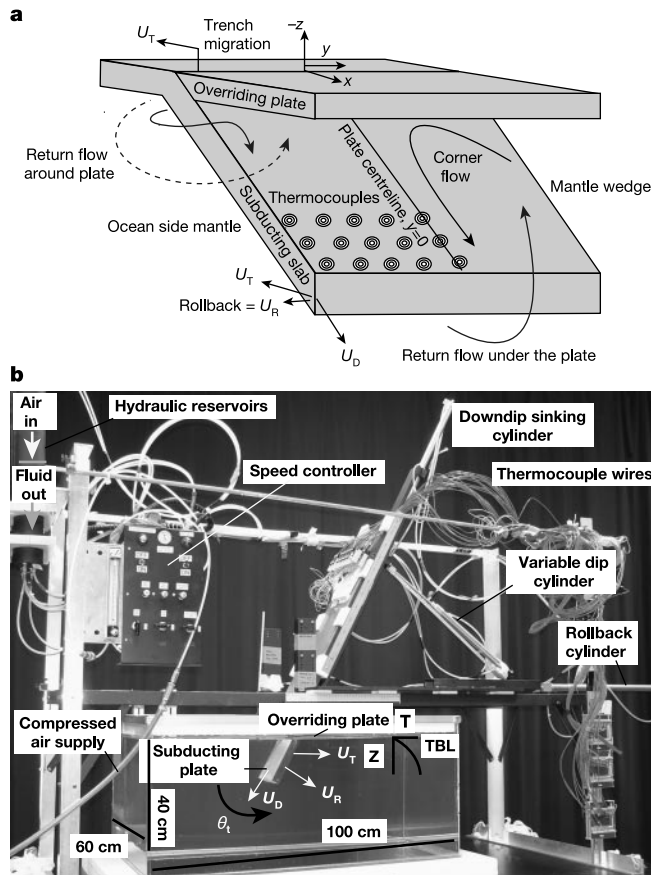


Figure 1 Subduction zone geometry and experimental set-up. **a**, Diagram illustrating different styles of plate sinking and mantle flow in subduction zones. Downdip sinking involves plate motion along a fixed trajectory at a rate (U_D) which drives a corner flow in the mantle wedge. Rollback subduction involves downdip and translational (U_T) plate velocities, and drives a mantle return flow from the ocean side of the plate to the wedge. The rollback velocity (U_R) is a vector sum of U_D and U_T . The coordinate axes are shown, with the position of the plate's axis of symmetry ($y = 0$). **b**, Photograph showing components of the plate subduction apparatus with the tank of glucose syrup that simulates the mantle (see Methods). T, temperature; Z, depth.

Table 1 Parameters for selected subduction laboratory experiments

Exp.	U_D	U_T	Dip	TBL	ΔT_5
4	2	0	49	0.7	1.4
7	2	1	49	0.7	-1.25
8	4.5	1	49	0.7	-0.1
9	2	2	49	0.7	-0.95
10	4.5	0	49	0.7	1.1
11	2	4	49	0.7	-0.6
12	2	0	74	0.7	0.75
13	4.5	0	74	0.7	0.65
14	11	0	49	0.7	0.3
15	11	0	74	0.7	0
16	2	4	74	0.7	-1.0
17	2	2	74	0.7	-1.1
19	2	0	θ_t	0.7	0.2
20	2	2	θ_t	0.7	-0.75
23	2	0	49	1.8	1.3
24	2	0	74	1.8	0.75
25	11	0	74	1.8	0.3
26	11	0	49	1.8	-0.4
28	2	2	49	1.8	-1.5
29	2	2	74	1.8	-1.6
30	2	2	θ_t	1.8	-1.6
31	2	0	θ_t	1.8	-1.3
32	2	0	49	0.7	1.3

Lateral temperature difference is calculated for thermocouple row 5 (ΔT_5) as edge minus centreline temperature. TBL is thermal boundary layer thickness defined as the depth in the fluid of the 15°C isotherm (scaled mantle value, $1,050^\circ \text{C}$). Assuming a $15\text{--}20\text{-km}$ -thick section of elastic plate, this scales to total lithospheric thicknesses for the overriding plate of $\sim 40 \text{ km}$ (for 0.7-cm cases) and $\sim 75 \text{ km}$ (for 1.8-cm cases). Dip angle is given in degrees from horizontal or as the parameter θ_t , which indicates variable dip cases where slabs steepen at $\theta_t = 2^\circ \text{ min}^{-1}$, from 49° to 74° . Lithosphere/ambient mantle viscosity contrast is infinite in the laboratory and $> 10^5$ in the mantle.

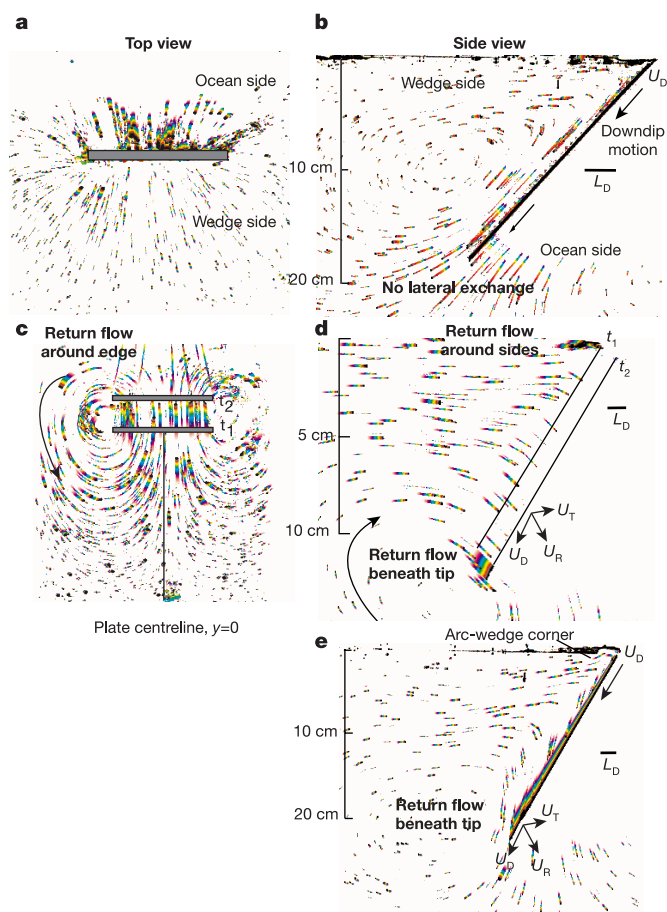


Figure 2 Streak patterns for tracers moving passively with the fluid over a time lapse interval (Δt_p) reveal circulation patterns (see Methods). The streak length (L) corresponding to plate motion is $L_D = U_D \Delta t_p$. Fluid velocity magnitude from individual streaks is then $L/L_D U_D$. Orientations include: **a**, a horizontal plane centred at 1.5 cm depth for a case of intermediate downdip sinking and a constant 49° dip angle for an exposure of $\Delta t_p = 1.4$ min. (exp. 32, Table 1). **b**, A vertical (x - z) section along the centreline of the fluid ($y = 0$) for the case in **a** and $\Delta t_p = 1.5$ min. The plate surface is marked as a dark line. **c**, A horizontal plane for a case of rollback subduction (exp. 30, Table 1). Positions of the plate are shown for start (t_1) and end (t_2) times for the streak image ($\Delta t_p = 2.5$ min). **d**, The vertical section through the plate's centre ($y = 0$) for this rollback case (**c**) is shown, with positions of the plate at t_1 and t_2 ($\Delta t_p = 1$ min). Regions in the wedge which are fed by flow around, versus flow beneath, the plate are indicated. At this instant the dip is 60° . Patterns shown in **c** and **d** are also similar to those for cases with similar U_D, U_T , but no steepening (exp. 29, Table 1). **e**, A vertical section (like **d**) for a case with intermediate downdip motion and plate steepening but with a fixed trench (exp. 31, Table 1) ($\Delta t_p = 1.2$ min). The region in the arc-wedge corner with higher velocities ($0.7 U_D$), along steeper trajectories is indicated.

Circulation in the wedge is very different for the case of rollback subduction (Fig. 2b, d), where the combination of corner flow driven by U_D and motion around the plate, due to U_T , produce a shear flow within the wedge (Fig. 3). Streak patterns in a shallow horizontal plane highlight the flow of material from the ocean to wedge side of the plate, in response to rollback motion (Fig. 2c). Velocities in the wedge are higher than the downdip case, particularly along the plate's centreline where material is preferentially drawn in towards the plate at speeds of 1.1 – $1.6 U_D$ (Figs 2c and 3). Flow towards the plate's upper surface is more sluggish nearer the edge ($y = W/2$) because the plate-ward velocity driven by U_D (Fig. 2a) is opposed by shearing stress due to flow from the ocean-side to the wedge-side of the plate (Figs 2b and 3).

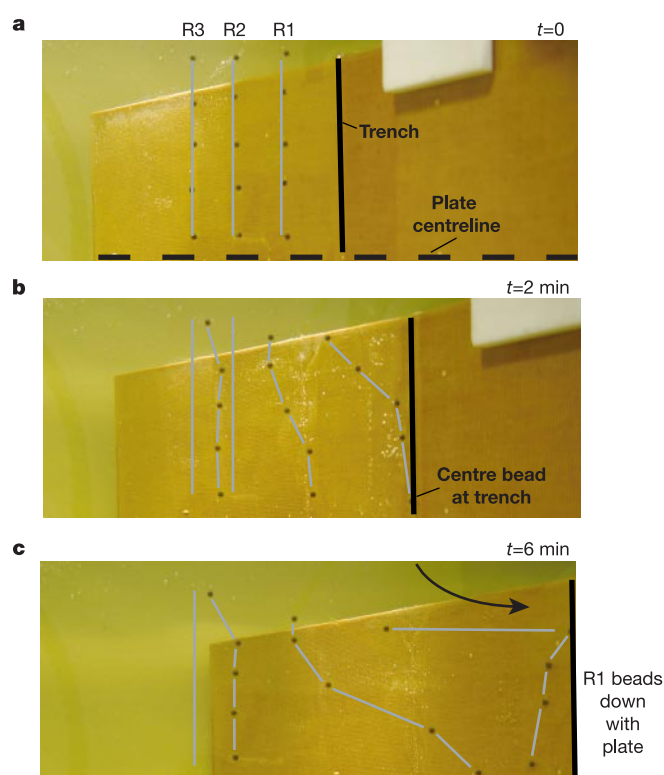


Figure 3 Displacement of passive Delrin beads in the shallow wedge owing to rollback. The view is looking down on the fluid, and highlights a strong shear flow. Plate rates are $U_D = U_T = 2 \text{ cm min}^{-1}$ and $\theta_t = 2^\circ \text{ min}^{-1}$, but for isothermal conditions with no overriding plate. **a**, Initial distribution of tracers is three rows of five beads, each at 1 cm depth. The fluid (plate) appears green (brown). **b**, After 2 min the beads are displaced by the rollback induced flow. Grey lines mark initial locations of rows. Beads from R1 show a strong lateral variation in flow towards the plate, with the centreline bead having reached the trench moving at 3.25 cm min^{-1} (or $1.6 U_D$). Beads from the edge of R1 and R2 move at 1 cm min^{-1} ($0.5 U_D$). **c**, After 6 min, three beads nearer the centreline from R1 have been subducted with the plate and R2 is sheared ($1.2 U_D$ for centreline bead ($y = 0$) versus $0.3 U_D$ for the bead at $y = W/2$). The dark arrow indicates the supply of material from around the plate edge.

The trajectories for the supply of material to the wedge vary in both space and time (Fig. 2d). During the early phase of subduction when the plate's leading edge is traversing the upper mantle, material from the ocean side of the plate moves under the plate's tip. As subduction progresses, only the deeper portion of the wedge is fed from beneath while the shallower region (< 7 – 8 cm) is supplied with material from around the edges of the plate (Fig. 2d). When the plate tip is deeper than 10 – 12 cm, the wedge is fed entirely by flow around the plate edges. A set of isothermal experiments with different plate widths show that this pattern in return flow around the plate edge is not sensitive to further increases in plate edge to sidewall spacing, although the magnitude of return velocities decreases with increasing spacing. Reducing the plate edge–sidewall spacing, which brings the flow closer to that of a 2D model approximation, is expected to modify this pattern of lateral flow by forcing more return flow beneath the plate tip.

There are striking differences in the thermal evolution of the wedge and plate between the two sinking styles. The rollback-induced focusing of the return flow (Figs 2c and 3) produces a 0.5 – 0.75°C lateral temperature variation in the wedge between the plate centre (warmer) and plate edge that is absent from the downdip cases. The effect is more pronounced in slab surface temperature (SST). For the cases of downdip subduction, the nearly

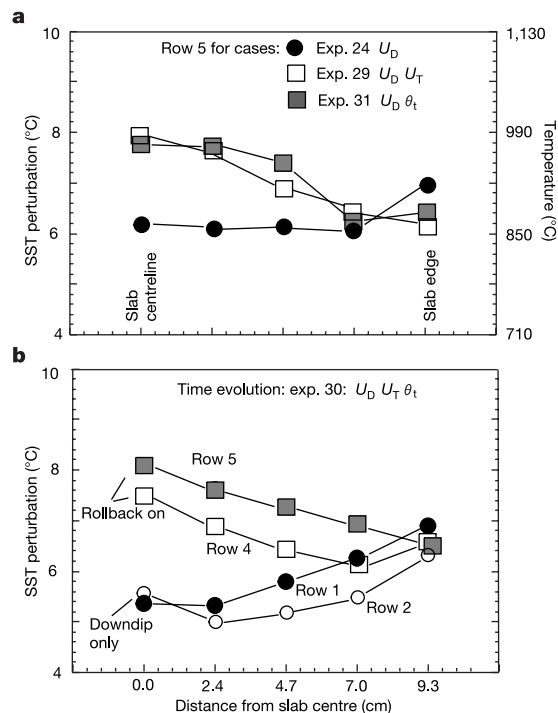


Figure 4 SST perturbation versus lateral distance across the plate. The perturbation is calculated by subtracting the initial SST (5 °C) from the value recorded at 5 cm depth (160 km). **a**, Plots compare data from thermocouple row 5 (near steady conditions) for cases with downdip motion along a fixed trajectory (exp. 24) and two modes of rollback motion (exp. 29, rollback translation without steepening; exp. 31, a fixed trench but with plate steepening). Higher temperatures are recorded at the edge for the case of purely downdip sinking, and at the plate centre for the rollback cases. The right hand y axis is the scaled mantle temperature (see Methods). **b**, Plots showing how lateral variation in SST evolves for the rollback case of Fig. 2c, d. Rows 1 and 2 pass through the system with only downdip motion. Once rollback is initiated, centreline temperatures quickly increase (rows 4 and 5).

uniform pattern of return flow towards the plate (Fig. 2a) is reflected in nearly constant temperatures across the central portion of the plate's upper surface (Fig. 4a). A robust feature of the constant-dip cases ($U_T = \theta_t = 0$) is that the plate edge follows a steeper trajectory through P - T space, where P is pressure and T is temperature. Temperature excesses (for example, edge-centre) can reach 1.4 °C by a depth of 5 cm, corresponding to 100 °C in the mantle (Fig. 4), and are enhanced for slower plate rates and shallower subduction angles (Table 1).

Rollback subduction produces a distinctly different along-arc SST distribution (Fig. 4). Warmer temperatures are consistently seen down the centreline of the plate, with a gradual decrease in SST towards the plate edge. Lateral temperature anomalies vary between 0.5 and 1.7 °C (mantle ~35–120 °C), and are enhanced for cases of subduction beneath a thicker TBL (1.8 cm cases in Table 1) and when plates sink with a component of slab steepening in addition to rollback motion (Fig. 4b). Higher temperatures near the centre of the plate are consistent with higher velocities in this region of the wedge. Stronger return flow in the wedge means fluid parcels lose less heat to the overlying plate and leads to a thinner TBL along the surface of the plate that enhances heat transfer into the plate.

Melting is not represented in these experiments, but the results have implications for melt production in arcs which may occur by decompression²⁷, melting of hydrated mantle and direct slab melting^{28,29} (for example, adakites). Plate melting is typically associated

with cases of slow subduction of young lithosphere^{8,9,28}, however it may also be an important feature of slab edges¹⁸. These results show that regions of elevated SSTs vary with position along the plate, and are sensitive to the style of slab sinking. Downdip sinking favours slab melting along plate edges, while rollback sinking favours slab melting along the centreline of the plate. Zones of active plate melting may alternate between the centre of the segment and the edges of the segment as the system oscillates between periods of downdip versus rollback subduction^{25,30}. A progression from rollback sinking to downdip sinking, or vice versa, produces a dramatic 3 °C swing in centreline SST at 5 cm depth (Fig. 4b) which is in excess of 200 °C at 160 km when scaled to the mantle. Future experiments will test whether asymmetric rollback (one end of the plate pinned) causes the region of elevated SST to migrate to the unpinned edge where the effects of higher translation rates and enhanced edge heating may combine.

Decompression melting in the mantle wedge²⁷ requires relatively warm temperatures and an upward component of motion. Downdip sinking on its own produces sluggish, nearly horizontal flow in the wedge. Rollback subduction results in faster wedge velocities that bring warmer material towards the centre of the plate, but flow trajectories tend to be horizontal. Return flow trajectories steepen within the arc-wedge corner when slab dip increases during sinking⁹ (Fig. 2e). On the ocean-side of the system, fluid in the vicinity of the plate edge moves horizontally around the edge and then downward with the plate (Fig. 1a). Further from the plate edge, material moves around the plate in a downward sweeping arc. After the plate passes, this material is drawn into the wedge and towards the plate along a slight upward trajectory (Fig. 1a), consistent with previous observationally based models^{15–17} (Fig. 2c). □

Methods

Modelling subducting and overriding plates

We model 3D flow associated with slab sections of finite length and as many as three components of slab motion. The subducting plate is a composite laminate of phenolic resin that is 20 cm wide and 2.5 cm thick. The scaled length (W) of the plate segment (along the trench) is 650 km, which is between values for the Scotia (450 km) and Marianas (1,200 km) plates. Three hydraulic piston cylinders for generating U_D , U_T and θ_t motions are driven with compressed air fed into a hydraulic reservoir, and stroke rates are controlled with precision flow meters in the hydraulic lines. Rollback occurs when all or part of the plate sinks with a component of translation (Fig. 1). Three rollback modes are: (1) $U_D + U_T$; (2) $U_D + \theta_t$; and (3) $U_D + U_T + \theta_t$. A 0.25-cm-thick plexiglas sheet overlies the wedge region of the fluid to simulate an overriding plate. For these cases, this plate migrates with the trench.

Initial conditions

Basic features of subduction zone thermal structures are simulated (cold plate; upper mantle TBL, shown schematically in Fig. 1b). Experiments begin with the plate apparatus coming to 5 °C in a constant-temperature room. The tank, containing fluid at a uniform 20 °C, is insulated (7.5-cm foam sheeting) on the sides and base and wheeled into the constant-temperature room. Subduction is initiated when the conductively growing surface TBL reaches the desired thickness (0.6 cm (1.8 cm) in 30 min (150 min)). The glucose viscosity varies exponentially with temperature following $\mu = 15 \exp[(1,800/(T + 93)) - 12.10]$. The maximum viscosity contrast within the glucose in these experiments is ~10, which is less than within the mantle. However, the very large viscosity increases within the sinking plate are represented with the elastic plate in the laboratory. The main difference then between laboratory and mantle viscosity structures occurs within a very narrow temperature range in the cool boundary layers beneath the overriding plate and around the sinking plate. This difference should not influence the basic patterns in flow or temperatures recorded for the different modes of plate sinking. To convert to mantle values (T_m) we use $T_m = T_{20} + [(T/15^\circ\text{C}) \times 1,050^\circ\text{C}] + (160 \text{ km} \times 0.5^\circ\text{C/km})$, where T_{20} is an estimate of mantle SST at 20 km depth and T are the SST perturbations shown here. Mantle values calculated using $T_{20} = 350^\circ\text{C}$ are shown in Fig. 4a.

Data acquisition

SSTs are monitored continuously with an array of thermocouples set into the plate (Fig. 1b), and wedge temperatures are recorded with descending probes at the completion of each experiment. Six rows of thermocouples (five per row) record plate surface temperatures (shown in Fig. 1a) at 1-s intervals. Distances of thermocouple rows from the slab tip are 1 cm (row 1), 8.5 cm (row 2), 23.5 cm (row 4) and 31 cm (row 5). Circulation patterns are monitored by illuminating the fluid with a thin 1.5-cm-wide sheet of light passed horizontally or vertically through the fluid, which reflects off microbubbles ($< 1 \text{ mm}$) suspended within the fluid (Fig. 2). Images are recorded over the entire

duration of each experiment with a digital video camera and processed using DigImage Software. In processing, there is a trade-off between particle density and streak length (product of velocity and time lapse interval). Streak images reveal an initial transient phase which rapidly evolves to near steady-state circulation. Therefore streak patterns in Fig. 2 represent near-steady pictures of wedge circulation. Estimates of fluid velocity can be made by comparing the lengths of individual streaks to that shown for the plate motion. The overlap of streaks in Fig. 2 occurs because of the time dependence of the rollback subduction, as viewed from the laboratory reference frame.

Received 29 November 2002; accepted 21 July 2003; doi:10.1038/nature01923.

- McCulloch, M. T. & Gamble, J. A. Geochemical and geodynamical constraints on subduction zone magmatism. *Earth Planet. Sci. Lett.* **102**, 358–374 (1991).
- Woodhead, J., Eggins, S. & Gamble, J. High field strength and transition element systematics in island arc and back-arc basin basalts: Evidence for multi-phase melt extraction and a depleted mantle wedge. *Earth Planet. Sci. Lett.* **114**, 491–504 (1993).
- Turner, S. & Hawkesworth, C. Constraints on flux rates and mantle dynamics beneath island arcs from Tonga-Kermadec lava geochemistry. *Nature* **389**, 568–573 (1997).
- Zhao, D., Hasegawa, A. & Horiuchi, S. Tomographic imaging of P and S wave velocity structure beneath Northeastern Japan. *J. Geophys. Res.* **97**, 19909–19928 (1992).
- Koper, K. D., Weins, D. A., Dorman, L. M., Hildebrand, J. A. & Webb, S. C. Constraints on the origin of slab and wedge anomalies in Tonga from the ratio of S to P anomalies. *J. Geophys. Res.* **104**, 15089–15104 (1999).
- Hsui, A. T., Marsh, B. D. & Toksoz, M. N. On melting of the subducted oceanic crust: Effects of subduction induced mantle flow. *Tectonophysics* **99**, 207–220 (1983).
- Davies, J. H. & Stevenson, D. J. Physical model of source region of subduction zone volcanics. *J. Geophys. Res.* **97**, 2037–2070 (1992).
- Peacock, S., Rushmer, T. & Thompson, A. Partial melting of subducting oceanic crust. *Earth Planet. Sci. Lett.* **121**, 227–244 (1994).
- Kincaid, C. & Sacks, I. S. The thermal and dynamical evolution of the upper mantle in subduction zones. *J. Geophys. Res.* **102**, 12295–12315 (1997).
- Kincaid, C. & Hall, P. S. The role of back-arc spreading in circulation and melting at subduction zones. *J. Geophys. Res.* **208**, 2240–2254 (2003).
- Elssasser, W. M. Sea floor spreading and thermal convection. *J. Geophys. Res.* **76**, 1101–1111 (1971).
- Klosko, E. *et al.* Upper mantle flow in the New Zealand region from seismic anisotropy. *Geophys. Res. Lett.* **26**, 1497–1500 (1999).
- Russo, R. & Silver, P. Cordillera formation, mantle dynamics and the Wilson cycle. *Geology* **24**, 511–514 (1996).
- Peyton, V. *et al.* Mantle flow at a slab edge: Seismic anisotropy in the Kamchatka region. *Geophys. Res. Lett.* **28**, 379–382 (2001).
- Pearce, J. A., Leat, P. T., Barker, P. F. & Millar, I. L. Geochemical tracing of Pacific-to-Atlantic upper-mantle flow through the Drake Passage. *Nature* **410**, 457–461 (2001).
- Turner, S. & Hawkesworth, C. Using geochemistry to map mantle flow beneath the Lau Basin. *Geology* **26**, 1019–1022 (1998).
- Wendt, J. I., Regelous, M., Collerson, K. D. & Ewart, A. Evidence for a contribution from two mantle plumes to the island-arc lavas from northern Tonga. *Geology* **25**, 611–614 (1997).
- Yogodzinski, G. M. *et al.* Geochemical evidence for the melting of subducting oceanic lithosphere at plate edges. *Nature* **409**, 500–504 (2001).
- Gvirtzman, Z. & Nur, A. The formation of Mount Etna as the consequence of slab rollback. *Nature* **401**, 782–785 (1999).
- Hudson, J. D. & Dennis, S. C. R. The flow of a viscous incompressible fluid past a normal flat plate at low and intermediate Reynolds numbers: The wake. *J. Fluid Mech.* **160**, 369–383 (1985).
- Garfunkel, Z., Anderson, C. A. & Schubert, G. Mantle circulation and the lateral migration of subducted slabs. *J. Geophys. Res.* **91**, 7205–7223 (1986).
- Ribe, N. Mantle flow induced by back-arc spreading. *Geophys. J. Int.* **98**, 85–91 (1989).
- Kincaid, C. & Olson, P. An experimental study of subduction and slab migration. *J. Geophys. Res.* **92**, 13832–13840 (1987).
- Guillou-Frotier, L., Buttles, J. & Olson, P. Laboratory experiments on the structure of subducted lithosphere. *Earth Planet. Sci. Lett.* **133**, 19–34 (1995).
- Griffiths, R. W., Hackney, R. I. & van der Hilst, R. D. A laboratory investigation of effects of trench migration on the descent of subducted slabs. *Earth Planet. Sci. Lett.* **133**, 1–17 (1995).
- Buttles, J. & Olson, P. A laboratory model of subduction zone anisotropy. *Earth Planet. Sci. Lett.* **164**, 245–262 (1998).
- Sisson, T. W. & Bronto, S. Evidence for pressure-release melting beneath magmatic arcs from basalt at Galunggung, Indonesia. *Nature* **391**, 883–886 (1998).
- Drummond, M. S. & Defant, M. J. A model for trondhjemite-tonalite-dacite genesis and crustal growth via slab melting: Archean to modern comparisons. *J. Geophys. Res.* **95**, 21503–21521 (1990).
- Sigmarsson, O., Martin, H. & Knowles, J. Melting of a subducting oceanic crust from U-Th disequilibria in austral Andean lavas. *Nature* **394**, 566–569 (1998).
- van der Hilst, R. & Seno, T. Effects of relative plate motion on the deep structure and penetration depth of slabs below the Izu-Bonin and Mariana island arcs. *Earth Planet. Sci. Lett.* **120**, 395–407 (1993).

Acknowledgements This work benefited from the Margins TEL, held in August 2000 in Eugene, Oregon. We thank the Equipment Development Laboratory at URI-GSO, and T. Beasley and C. Morgan for technical assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.K. (kincaid@gso.uri.edu).

Craniometric evidence for Palaeoamerican survival in Baja California

Rolando González-José¹, Antonio González-Martín², Miquel Hernández¹, Héctor M. Pucciarelli³, Marina Sardi³, Alfonso Rosales⁴ & Silvina Van der Molen⁵

¹Universitat de Barcelona, Facultat de Biologia, Secció d'Antropologia, Diagonal 645, 08028 Barcelona, Spain

²Universidad Autónoma del Estado de Hidalgo, Área Académica de Historia y Antropología, 42000 Pachuca, México

³Departamento Científico de Antropología del Museo de La Plata, Facultad de Ciencias Naturales y Museo, Universidad Nacional de La Plata, 1900 La Plata, Argentina

⁴Instituto Nacional de Antropología e Historia, Centro INAH Baja California Sur, 23000 La Paz, México

⁵Universitat Autònoma de Barcelona, Facultat de Ciències, BAVE, Unitat de Zoologia 08193, Bellaterra, Barcelona, Spain

A current issue on the settlement of the Americas refers to the lack of morphological affinities between early Holocene human remains (Palaeoamericans) and modern Amerindian groups, as well as the degree of contribution of the former to the gene pool of the latter^{1–6}. A different origin for Palaeoamericans and Amerindians is invoked to explain such a phenomenon³. Under this hypothesis, the origin of Palaeoamericans must be traced back to a common ancestor for Palaeoamericans and Australians, which departed from somewhere in southern Asia and arrived in the Australian continent and the Americas around 40,000 and 12,000 years before present, respectively. Most modern Amerindians are believed to be part of a second, morphologically differentiated migration³. Here we present evidence of a modern Amerindian group from the Baja California Peninsula in Mexico, showing clearer affinities with Palaeoamerican remains than with modern Amerindians. Climatic changes during the Middle Holocene probably generated the conditions for isolation from the continent, restricting the gene flow of the original group with northern populations, which resulted in the temporal continuity of the Palaeoamerican morphological pattern to the present.

A study by Neves *et al.*³ first suggested that the pattern of craniofacial affinities of humans in the Americas could be explained by a double migratory event: a first migration led by Palaeoamericans originating from an ancestral population inhabiting Asia in pre-glacial times, and a second migration of the so called Mongoloids from which derived most of the modern Amerindians. If this hypothesis is correct then we should observe relicts of the former Palaeoamerican stock somewhere in the New World, especially in geographically isolated areas where the lack of gene flow could have enabled the persistence of genetic and phenotypic ancestral traits.

Thirty-three skulls stored at the Regional Museum of La Paz and National Museum of Anthropology and History (México) were analysed and measured. Remains were assigned to the ethnographical Pericú group, which occupied the southern tip of the peninsula of Baja California Sur (Mexico) and displayed a strict hunter-gatherer subsistence pattern⁷. Skulls belong to Late Holocene horizons and were recovered in contextualized archaeological excavations⁸ (see Methods for more details). Twenty-four Howells variables⁹ were measured on each cranium from the Baja California series (BCS), and on a set of samples including the Palaeoamerican series of Lagoa Santa^{1–6} (Table 1). These craniometric traits were used to compute unbiased minimum genetic distances by means of

the Relethford–Blangero method^{10–13}. Furthermore, a canonical variates analysis was carried out to assess the affinities of samples and to evaluate the relative variability within the samples.

Geometric morphometrics were used to visualize shape changes as landmark deformation, and to perform multivariate analysis on landmark coordinates (see Methods). Major shape changes in projected lateral view were shown using thin-plate spline¹⁴, after removing scale, position and rotation by generalized Procrustes superimposition^{15,16}. Additionally, Goodall's *F*-test was computed to test for overall shape differences between groups^{15,17}. This statistic is based on the ratio of the squared Procrustes distance between the means of each sample, to the sum of the squared Procrustes distance of each specimen to its group mean. Furthermore, an evaluation of shape change based on relative warp analysis is provided as Supplementary Information.

The principal coordinate plot obtained using the matrix of minimum genetic distances (Fig. 1a) showed that BCS was closely linked with Palaeoamericans, whereas the populations from Patagonia and Tierra del Fuego showed an intermediate position between classical Amerindians and/or East Asians and Palaeoamericans^{5,18}. Even when SANT (California) is the nearest group to BCS in geographical and chronological terms, they did not share clear craniofacial affinities, with SANT being more similar to PERU, ARIK, BOL, TOBA and Asian samples (see Table 1 for definitions). The results of the canonical variates analysis (Fig. 1b) are coincident with the Relethford–Blangero distances, and show clearly the overlapping of Baja Californian specimens within the Palaeoamerican spectrum of variability, as well as their departure from the variability accounted for the remaining Amerindians.

Shape differences between three samples (BCS, TLAT and four selected Palaeoamericans) explored by geometric morphometrics are presented in Fig. 2. Results highlight the similarities between facial and neural projected shapes of BCS and Palaeoamericans in contrast with comparisons between TLAT and Palaeoamericans. Thin-plate spline analyses reflect differences between Palaeoamericans and BCS mainly due to a slight shortening of the neurocranium enclosed with a diminution of the glabellar region and a

slight increase in prognathism. Goodall's *F*-test revealed that these differences are not significant. Conversely, comparison between TLAT and Palaeoamericans as well as TLAT and BCS showed changes consistent with an important shortening of the neurocranium, mainly concentrated in the opisthocranium, and an expansion of the facial cranium of TLAT in relation to BCS and Palaeoamericans. In both comparisons, differences were highly significant. The most obvious differences between TLAT and BCS or Palaeoamericans are that faces of the latter group are much smaller relative to overall cranial size, and the neurocranium is more elongated than in TLAT.

In general terms, similarities in craniofacial patterns can be explained by three different factors: gene flow, convergent adaptation to local environment and common ancestry combined with isolation. Gene flow can be discarded here because BCS and Palaeoamericans are separated by significant geographical and

Table 1 Populations studied, codes and sample sizes

Population and region	Reference	Code	<i>n</i>
Baja California Sur, Mexico	*	BCS	33
Fueguians, Tierra del Fuego	5	FUEG	89
Patagonians, Patagonia	5	PATA	98
Andean Patagonians, Patagonia	5	APAT	20
Pampas, Buenos Aires, Argentina	5	PAM	23
Delta of Paraná, eastern Argentina	*	DPAR	30
Aztecs, Mexico	*	TLAT	39
Bolivians, Bolivia	†	BOL	19
Toba, northeast Argentina	†	TOBA	10
Calchaquí, northwest Argentina	†	CAL	8
Palaeoamericans of Brazil	1–5	PAL	22
Teita, Kenya	9	TEITA	83
Dogon, Mali	9	DOGON	99
Zulu, South Africa	9	ZULU	101
Bushman, Kenya	9	BUSH	90
Australians, Australia	9	AUST	101
Tasmanians, Tasmania	9	TASM	87
Tolai, Melanesia	9	TOLAI	110
Buriats, East Asia	9	BURIAT	109
Eskimo, Greenland	9	ESKI	108
Yauyos, Peru	9	PERU	110
Arikara, USA	9	ARIK	69
Ainu, Japan	30	AINU	86
North Japanese, Japan	30	NJAP	87
South Japanese, Japan	30	SJAP	91
Hainan, southern China	30	HAIR	83
Anyang, Taiwan	30	ANYA	42
Atayal, eastern China	30	ATAY	47
Santa Cruz, California, USA	30	SANT	102

* Present study.

† Present study, measured by M.S.

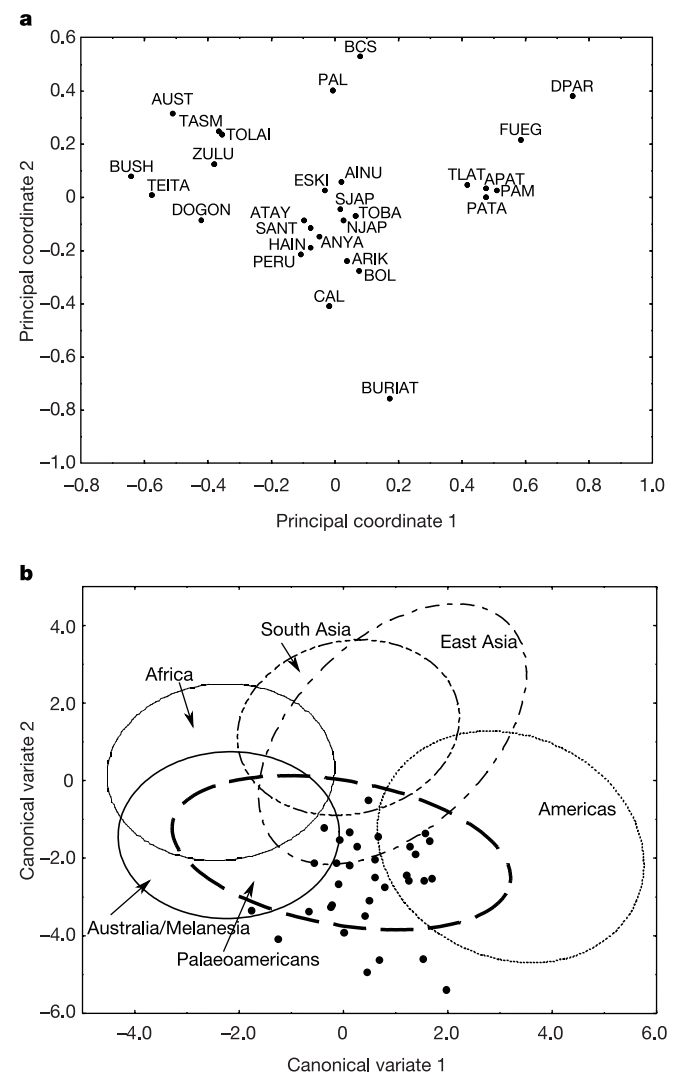


Figure 1 Results concerning the multivariate analysis of Howells variables. **a**, Principal coordinates plot representing the matrix of unbiased minimum genetic distances obtained after the Howells variables and computed according to references 10–13. Sex was estimated following the standard criteria, and data were z-standardized to avoid sex-related scaling effects¹². Heritability was set to a value of 0.55 (refs 2, 11), and equal population sizes were assumed. Unbiased minimum genetic distances were obtained after the correction for sampling bias¹³. **b**, Canonical variates analysis on the z-standardized raw data. Ninety per cent equal frequency ellipses were drawn for the first two canonical variates. Individual Californians (filled circles) are plotted.

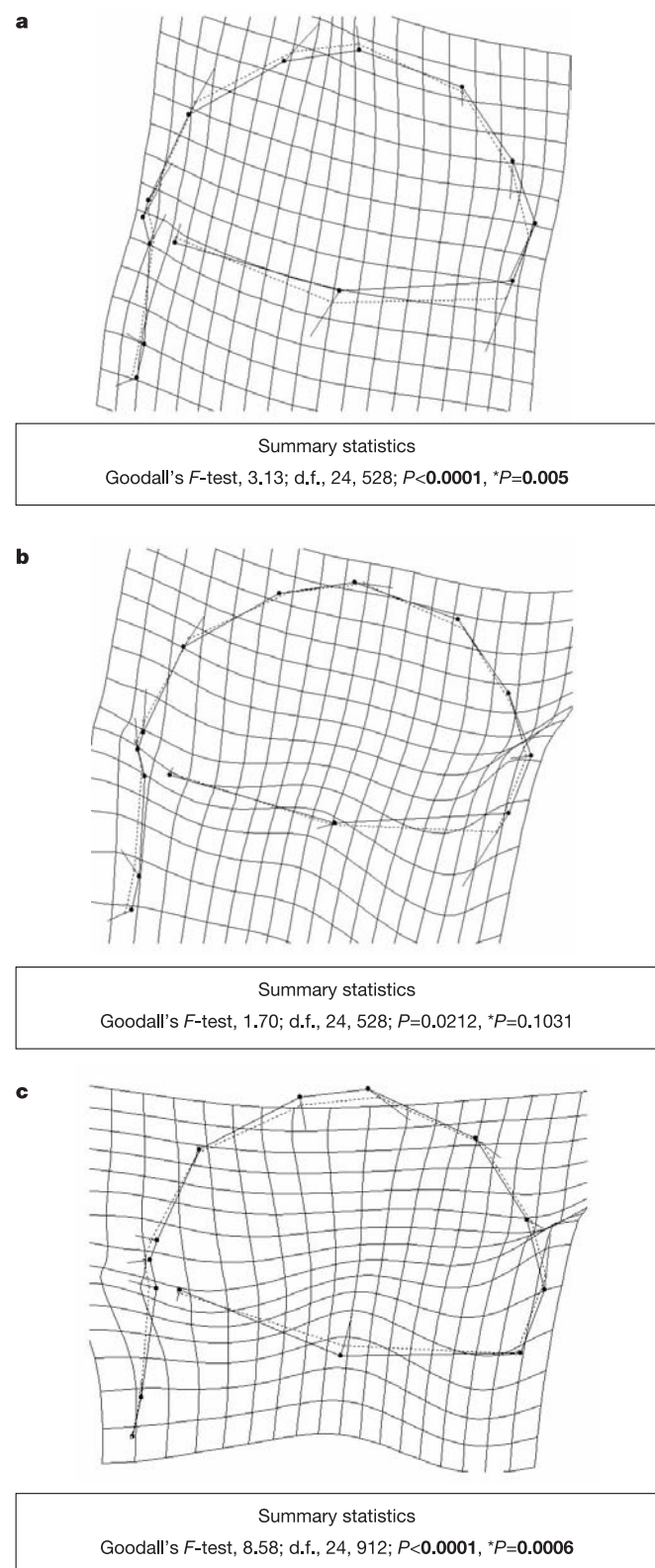


Figure 2 Geometric morphometric comparisons of BCS, TLAT and Palaeoamericans. **a–c**, Thin-plate spline was based on generalized Procrustes superimposition of TLAT (target) and Palaeoamericans (reference) (**a**), BCS (target) and Palaeoamericans (reference) (**b**) and BCS (target) and TLAT (reference) (**c**). References are represented by filled circles and solid lines; targets are displayed as open circles and dotted lines. Vectors and spline are magnified by $\times 4$. Boxes contain the Goodall's F -test statistics for the paired groups, degrees of freedom (d.f.), associated P -values, and the permutational P -values¹⁷ (marked with asterisks). Significant values are in bold.

temporal distances. The position of some groups in Fig. 1a (especially the Buriats) has led to the statement that the divergent nature of these groups might reflect environmental variation, specifically variation in overall cranial size due to climate. However, such divergence levels in human populations can also be generated by other microevolutionary mechanisms such as genetic drift and isolation. In the case of Buriats, isolation might be an important source of divergence. For the moment, there is no clear empirical or theoretical evidence pointing to environmental adaptation as a main factor shaping craniofacial morphology in *Homo sapiens*, except for some gracilization as a response to reduced masticatory stress¹⁸. This does not imply that the environment did not affect at all some particular cranial phenotypic traits, but indicates that, after computing morphological distances, its effect is rather small^{19–21}. Furthermore, an inspection of palaeoclimatic studies of Baja California and Amazonia demonstrates that basic climatic and environmental parameters remained significantly divergent for both regions since the Late Pleistocene. Palynological and stratigraphical records reveal that wet phases on both the Amazonian basin and the Sonoran desert are correlated, but never derived in similar environmental conditions after the arrival of the first humans^{22–24}.

Even though the results presented here cannot solve the issue of environment on craniofacial morphology, an environmentally led convergence of craniofacial traits in both populations seems unlikely because climatic conditions differed during the whole period of human settlement. Although we cannot completely discard micro-environmental adaptation as a possible factor, the results presented here could be better explained by a history of isolation between early Baja California inhabitants and their descendants, and the latter Amerindian groups inhabiting mainland areas. Geographical isolation could be evoked by the appearance, in the Middle Holocene, of the Sonoran desert which 'cut in two' the Baja California Peninsula. This process may have had an effect on the evolution, distribution and genetic structure of Baja California's terrestrial vertebrates^{25,26}. Therefore it is likely that human populations arriving at the southern tip of the peninsula through demographic pressure exerted by the northern groups also suffered some degree of barrier to gene flow throughout the Sonoran desert. Such genetic isolation was possible even when cultural mobility might have been intense during the Early Holocene. The southward region probably acted as a refuge in which human groups remained spatially, genetically and perhaps culturally isolated from the mainland. In this model, the inferred presence of Palaeoamericans in the southern tip of the peninsula supports the hypothesis of a "Pacific route" of settlement²⁷. More hypothetical population relicts or admixture areas must be explored in the future to interpret the population dynamics and genetics of the first Americans. □

Methods

The Baja California sample

Pericú is included in the Guaicurian language family, which is different from the Yuman spoken in the centre and north of the peninsula. However it may possibly be related to Yuman in the larger Hokan linguistic group. Culturally, the Guaicurian tribes differed noticeably from the more distant Yumans in the north, as exhibited in the ethnographic horizon⁷. The extinction of these groups was abruptly provoked by disassembling of the hunter-gather system due to the influence of Jesuitical missions. The materials analysed came from caves and coastal shell middens. Primary and secondary burials were common at both sites, and these mortuary practices are typical of the late prehispanic cultures, which led to the ethnohistoric cultures. Lithic technology is clearly adapted to the exploitation and processing of marine resources, and sites usually lack a space allocated for the elaboration of tools⁸. See Supplementary Information for further details on this issue.

Geometric morphometrics

Landmark coordinate data were available for a subsample of 20 Mexican Amerindians (Aztecs from Tlatelolco), 20 specimens from Baja California Sur, and four Palaeoamerican skulls from Mexico and Brazil. Fourteen two-dimensional landmarks were taken directly from digitized images. Goodall's F -test requires the assumption that landmarks are symmetrically distributed around the Procrustes mean (isotropy). A visual inspection of the two-dimensional scatter plots, and a principal component analysis on the shape

dimensions matrix of these data sets¹⁷ indicates that this is a reasonable assumption. Twenty individuals were selected for each BCS and TLAT digitized image in order to achieve sex-balanced samples. The Palaeoamerican series contains four individuals: two males and two females (Lapa Vermelha IV, Peñon III, Chimalhuacán and Metro Balderas; refs 3, 6). The landmarks used were prosthion, anterior subnasal spine, nasion, glabella, most inferoposterior midline point above glabella (frontex), midline point of greatest elevation between nasion and bregma (metopion), bregma, vertex, midline point of greatest elevation between vertex and lambda, lambda, opisthocranium, inion, porion and maxillofrontal point. Thin-plate spline and Goodall's *F* tests were computed using the Morphue package²⁸ and the TwoGroup subroutine of the IMP package²⁹ respectively. We used the Bonferroni method to adjust for the lack of independence in the comparisons between populations. The total comparisons were $c = k(k - 1)/2$ (where k is the number of populations), and the significance level 0.05 was divided by c (P -value = $0.05/3 = 0.0166$) to guarantee the real significance.

Received 19 February; accepted 19 May 2003; doi:10.1038/nature01816.

1. Neves, W. A. & Pucciarelli, H. M. Morphological affinities of the first Americans: an exploratory analysis based on early South American human remains. *J. Hum. Evol.* **21**, 261–273 (1991).
2. Powell, J. F. & Neves, W. A. Craniofacial morphology of the first Americans: pattern and process in the peopling of the New World. *Yearb. Phys. Anthropol.* **42**, 153–188 (1999).
3. Neves, W. A., Powell, J. F. & Ozolins, E. G. Extra-continental morphological affinities of Lapa Vermelha IV, Hominid 1: A multivariate analysis with progressive numbers of variables. *Homo* **50**, 263–282 (1999).
4. Neves, W. A., Powell, J. F. & Ozolins, E. G. Extra-continental morphological affinities of Palli-Aike, Southern Chile. *Interciencia* **24**, 258–263 (1999).
5. González-José, R., Dahinten, S. L., Luis, M. A., Hernández, M. & Pucciarelli, H. M. Craniometric variation and the settlement of the Americas: testing hypotheses by means of R matrix and matrix permutation tests. *Am. J. Phys. Anthropol.* **116**, 154–166 (2001).
6. González-José, R., Neves, W., Hernández, M., Pucciarelli, H. & Correal, C. Late Pleistocene/Early Holocene skulls from Mexico demonstrate the existence of the Paleoamerican morphological pattern in Mesoamerica. (submitted).
7. Willey, G. R. *An Introduction to American Archaeology* (Prentice-Hall, New Jersey, 1966).
8. Rosales-López, A. & Fujita, H. *La Antigua California Prehispánica: la Vida Costera en El Conchalito* (INAH, México, 2000).
9. Howells, W. W. *Cranial Variation in Man* (Papers of the Peabody Museum, Harvard University, Massachusetts, 1973).
10. Relethford, J. H. & Blangero, J. Detection of differential gene flow from patterns of quantitative variation. *Hum. Biol.* **62**, 5–25 (1990).
11. Relethford, J. H. & Harpending, H. C. Craniometric variation, genetic theory and modern human origins. *Am. J. Phys. Anthropol.* **95**, 249–270 (1994).
12. Williams-Blangero, S. & Blangero, J. Anthropometric variation and the genetic structure of the Jirels of Nepal. *Hum. Biol.* **61**, 1–12 (1989).
13. Relethford, J. H., Crawford, M. H. & Blangero, J. Genetic drift and gene-flow in post famine Ireland. *Hum. Biol.* **69**, 443–465 (1997).
14. Bookstein, F. L. Principal warps—Thin-Plate Splines and the decomposition of deformations. *Trans. Pat. Anal. Mac. Intel.* **11**, 567–585 (1989).
15. Goodall, C. R. Procrustes methods in the statistical analysis of shape. *J. R. Stat. Soc. B* **53**, 285–339 (1991).
16. Rohlf, F. J. Shape statistics: Procrustes superimpositions and tangent spaces. *J. Classif.* **16**, 197–223 (1999).
17. Dryden, I. L. & Mardia, K. V. *Statistical Shape Analysis* (Wiley, Chichester, 1998).
18. Lahr, M. M. *The Evolution of Modern Human Diversity: A Study of Cranial Variation* (Cambridge Univ. Press, Cambridge, 1996).
19. Relethford, J. H. Apportionment of global human genetic diversity based on craniometrics and skin color. *Am. J. Phys. Anthropol.* **118**, 393–398 (2002).
20. Sparks, C. S. & Jantz, R. L. A reassessment of human cranial plasticity: Boas revisited. *Proc. Natl Acad. Sci. USA* **99**, 14636–14639 (2002).
21. González-José, R., Van der Molen, S., González-Pérez, E. & Hernández, M. Patterns of phenotypic covariation and correlation in modern humans as viewed from morphological integration. *Am. J. Phys. Anthropol.* (in the press).
22. Bradley, R. S. *Paleoclimatology: Reconstructing Climates of the Quaternary* (Academic, San Diego, 1999).
23. Haberle, S. Late Quaternary vegetation and climate history of the Amazon Basin: correlating marine and terrestrial pollen records. *Proc. Ocean Drilling Prog. Sci. Res.* **155**, 381–396 (1997).
24. Baker, P. A. *et al.* The history of South American Tropical precipitation for the past 25,000 years. *Science* **291**, 640–643 (2001).
25. Grismer, L. Evolutionary biogeography on Mexico's Baja California peninsula: a synthesis of molecules and historical geology. *Proc. Natl Acad. Sci. USA* **97**, 14017–14018 (2000).
26. Riddle, B. R., Hafner, D. J., Alexander, L. F. & Jaeger, J. R. Cryptic vicariance in the historical assembly of a Baja California Peninsular desert biota. *Proc. Natl Acad. Sci. USA* **97**, 14438–14443 (2000).
27. Dixon, E. J. Human colonization of the Americas: timing, chronology and process. *Quat. Sci. Rev.* **20**, 277–299 (2001).
28. Slice, D. E. Morphue *et al.* software for morphometric research. Revision 01-30-98 (Department of Ecology and Evolution, State University of New York, Stony Brook, 1998).
29. Sheets, H. D. Integrated Morphometrics Package (Department of Geology, SUNY at Buffalo, New York, 2001).
30. Howells, W. W. *Skull Shapes and the Map* (Papers of the Peabody Museum, Harvard University, Massachusetts, 1989).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We would like to thank personnel and authorities of the Instituto Nacional de Antropología e Historia, Museo Nacional de Antropología e Historia (México), and Musée de l'Homme (France) for their collaboration during the data acquisition. We thank P. Nepomnaschy, C. O'Connor, and J. Garrett for comments on the manuscript. We are also indebted to E. Ozolins who made available his data on Palaeoamerican series. This work was partially funded by the Wenner-Gren Foundation for Anthropological Research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.G.J. (rolandogonzalez@ub.edu).

A uniquely specialized ear in a very early tetrapod

J. A. Clack¹, P. E. Ahlberg^{2*}, S. M. Finney¹, P. Dominguez Alonso², J. Robinson² & R. A. Ketcham³

¹University Museum of Zoology, Cambridge, Downing Street, Cambridge, CB2 3EJ, UK

²Department of Palaeontology, Natural History Museum, Cromwell Road, London SW7 5BD, UK

³High-resolution X-ray CT Facility, Department of Geological Sciences, 1, University Station C1110, University of Texas, Austin, Texas 78712, USA

* Present address: Department of Evolutionary Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18A, SE-752 36 Uppsala, Sweden

The Late Devonian genus *Ichthyostega* was for many decades the earliest known tetrapod, and the sole representative of a transitional form between a fish and a land vertebrate. However, despite being known since 1932 (ref. 1) from a large collection of specimens, its morphology remained enigmatic and not what was expected of a very primitive tetrapod². Its apparent specializations led it to be considered as a “blind offshoot”³ or “side-branch”⁴ off the tetrapod family tree, and recent cladistic analyses have disagreed about its exact phylogenetic position^{5–8} within the tetrapod stem group. In particular, its braincase and ear region defied interpretation, such that conventional anatomical terms seemed inapplicable⁴. Using new material collected in 1998 (ref. 9), preparation of earlier-collected material, and high-resolution computed tomography scanning, here we identify and interpret these problematic anatomical structures. They can now be seen to form part of a highly specialized ear, probably a hearing device for use in water. This represents a structurally and functionally unique modification of the tetrapod otic region, unlike anything seen in subsequent tetrapod evolution. The presence of deeply grooved gill bars as in its contemporary *Acanthostega*¹⁰ suggest that *Ichthyostega* may have been more aquatically adapted than previously believed.

Three specimens have contributed most to the reinterpretation of the braincase region of *Ichthyostega*. MGUH (Geological Museum, Copenhagen) 6018 shows the posterior part of the braincase and stapes in ventral view (Fig. 1a, b). MGUH field number (f.n.) 200 shows the otoccipital region in dorsal view and branchial elements (Fig. 1c, d). MGUH f.n. 180 comprises the right half of a three-dimensionally-preserved skull roof and braincase (Fig. 1e, f), of which the posterior part has been scanned by computed tomography (CT). Figure 2 shows three-dimensional models and an interpretive diagram of part of MGUH f.n. 180 (see Methods and Supplementary Movies).

The new information reveals a highly specialized otoccipital configuration in *Ichthyostega* (Fig. 3a, c, e, f). It resembles neither those of tetrapodomorph fish such as *Eusthenopteron*³ and *Pander-*

*ichthys*¹¹, nor those of other early tetrapods such as *Acanthostega*¹² (Figs 3b, d, e and 4) or later forms^{13,14}. The inner ear sacculus chambers, clearly visible in the CT scans, are remarkably small and anteriorly placed (Fig. 2f). Semicircular canal tracts have not been certainly identified because of the limited resolution of the CT scans and the poor internal ossification of the otic capsules, but the lateral canal may lie in a convexity seen in dorsal view (Figs 1d and 2a, c, f). It appears to have been correspondingly small. The posterior part of the braincase housing the hindbrain is by contrast, extremely elongated (Figs 1a, b, 2a, c and 3a, c). Early tetrapods generally have smaller inner ears than similarly sized sarcopterygian fishes¹⁵, but the combination of a tiny ear and long hindbrain in *Ichthyostega* is unique. The otoccipital forms a narrow crest that sutures dorsally to the skull roof (Figs 1d, 2a and 3c), underneath the single fused postparietal that is also unique to *Ichthyostega*. The opisthotics form large posterior flanges suturing to down-turned extensions of the skull table margin, and anteriorly the proötics are developed into strong transverse buttresses (Figs 1a–d, 2b and 3a, c). The structure contrasts with the rather short, dorsally flat and more or less square otic region of tetrapods such as *Acanthostega* and many Carboniferous forms^{7,12–14} (Fig. 3b, d).

On either side of the otoccipital braincase block is a large chamber, defined by solid walls formed of several different but closely sutured bones (proötic, opisthotic, epipterygoid and skull table; Figs 2b and 3a, c). The chamber extends above the otic capsules to the midline crest. This 'otic chamber' is unique to *Ichthyostega* (Fig. 3a, f); in other early tetrapods the skull table margin at this point is simple and narrow. Projecting into the chamber from the undersurface of the skull table is a finger-like scrolled structure named 'Säve-Söderbergh's process'^{2,4} (Figs 1a, b, 2b and 3a), again seen nowhere else in early tetrapods. The position of the otic (or spiracular) notch is similarly different from other early tetrapods (Fig. 3e) and panderichthyid fishes¹¹, being more posterior and facing laterally.

The stapes itself is unlike any other known tetrapod stapes or any fish hyomandibula. Previously misidentified as the 'sacculus vesicle'⁴, its 'shaft' is a very thin, almost circular, anterodorsally curved lamina of bone, projecting dorsolaterally into the otic chamber. (Figs 1 and 2). It contrasts with the short, stout, robust stapes with a single large head that is now recognized in many other early tetrapods¹⁶ (Figs 3b and 4). The stapes of *Ichthyostega* appears to be double headed (the two heads just meet but do not appear to

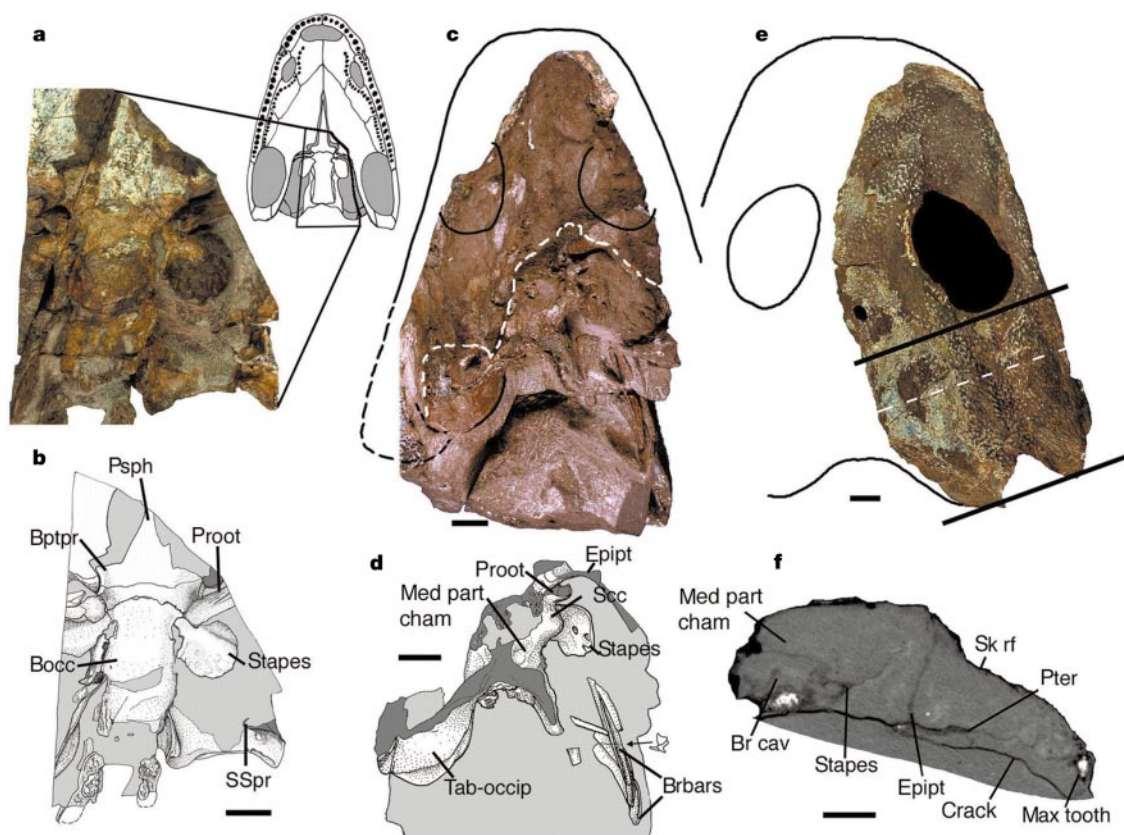


Figure 1 *Ichthyostega* sp. specimens used for the study. **a**, MGUH 6018, collected 1932 from north Celsius Bjerg but previously undescribed, mechanically prepared in ventral view to show the elongated basioccipital and the almost circular stapes that approaches very closely to the proötic buttress. Säve-Söderbergh's process is seen on the right (morphological left). Inset, diagram of the palate to show the region represented in the specimen. **b**, Interpretive drawing of MGUH 6018. **c**, Part of MUGH f.n. 200, collected 1998 from south Celsius Bjerg, prepared in dorsal view to show a horizontal section through the otoccipital, the stapes, part of the flanges closing off the otic chamber posteriorly, and a set of grooved branchial elements. Curved line, extrapolated skull margin; white dashed line, anterior extent of drawing shown in **d**. **d**, Interpretive diagram

of MGUH f.n. 200. The inset to right of arrow shows cross-section through the gill bars. **e**, MGUH f.n. 180, collected 1998 from south Celsius Bjerg, the right half of an almost uncrushed skull. Black lines, the limits of the CT scan series and approximate skull outline; white dashed line, approximate position of slice shown in **f**. **f**, One of the CT scan sections through MGUH f.n. 180. In the diagrams, light grey areas indicate matrix and dark grey areas indicate bone. Scale bars, 10 mm. Bocc, basioccipital; Bptpr, basiptyergoid process; Br cav, brain cavity; Brbars, branchial bars; Epipt, epipterygoid; Scc, semicircular canal; Max, maxilla; Med part cham, medial part of the otic chamber; Proot, proötic buttress; Pter, pterygoid; Sk rf, skull roof; SSpr, Säve-Söderbergh's process; Tab-occip, tabular-occipital complex.

form a true footplate), with a large stapedal foramen separating them (Fig. 2d, e). The ventral head contacts the basioccipital at a small facet, whereas the dorsal head inserts into the fenestra vestibuli (or fenestra ovalis), which communicates directly with the saccular chamber. Vertical movement of the shaft would thus have caused a piston-like in/out movement of the dorsal head in the fenestra, well suited to transmitting vibrations into the inner ear (Fig. 3f).

The otic chamber cannot be equivalent to the post-temporal fossa, housing the epaxial musculature, of other tetrapodomorphs, because it is closed off posteriorly by the down-turned skull table margin. It is also entirely separated from the jaw adductor musculature by the epipterygoid (Fig. 3f). The position of the chamber relative to known anatomical landmarks leaves little doubt that it housed the spiracular tract (or middle ear cavity), which must thus have been extremely specialized and greatly enlarged relative to those of other early tetrapods and sarcopterygian fishes. The position and orientation of the stapes within the chamber suggests the separation of the chamber into different regions, one of which extended between the skull roof and the dorsal surface of the stapes and may have been demarcated posteriorly by Säve-Söderbergh's process.

This structure is not that of a tympanic ear. The spiracular notch is in quite a different position from that in tetrapods with tympanic

ears such as temnospondyls, the stapes is in an inappropriate position, is orientated away from the notch, and is quite different in shape from those of tympanic ears. As for the 'middle ear' (or spiracular tract) itself, it could in principle have been occupied by either water or air. The phylogenetic position of *Ichthyostega* within the lungfish–tetrapod clade suggests that it was at least a facultative air-breather, implying that the spiracular tract could easily have been air-filled even if the animal was essentially (and primitively) aquatic. Functionally, the most plausible interpretation of the ear seems to be that the chamber housed an enlarged air pocket (Fig. 3f). The stapes was not in contact with any hard-tissue structure except for its delicate articulation with the braincase. We believe it may have been embedded in a soft-tissue membrane forming the ventral wall of the otic chamber. Notably, the space between the anterior margin of the stapes and the laterally projecting prootic buttress (Figs 1a, b, 2a and 3a), which on comparative anatomy criteria should connect the eustachian tube with the dorsal part of the spiracular tract and thus should not be bridged by a membrane, is very narrow in *Ichthyostega* compared with other early tetrapods. This suggests that the passage was strongly constricted, functionally sealing off the dorsal air chamber as a closed unit. Säve-Söderbergh's process may have supported a muscularized valve to seal off the chamber posteriorly. Reinforcement of the dorsal and marginal walls of the chamber to form a rigid bony box suggests a device to

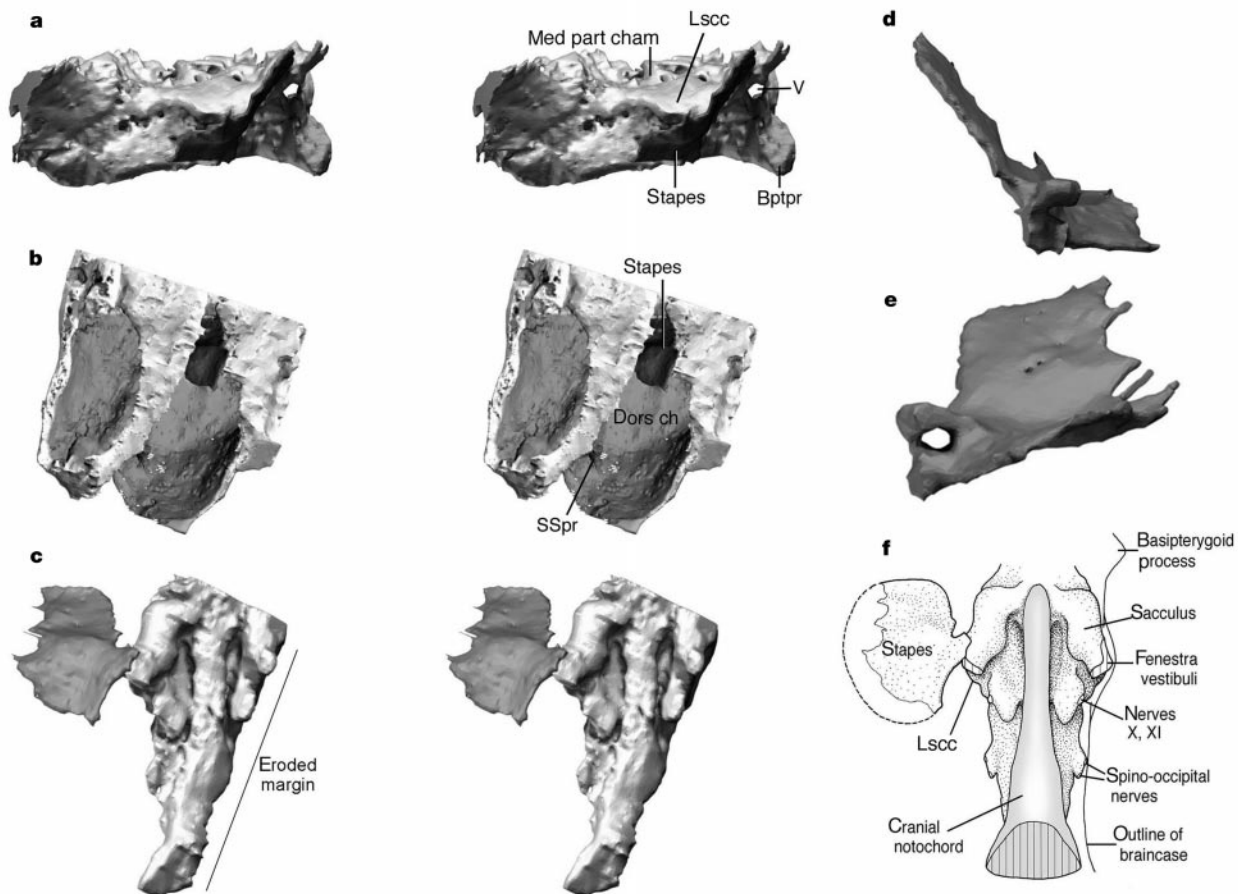


Figure 2 Models created from CT scan series of MGUH f.n. 180. **a**, Stereo pair of the right lateral view of the braincase with the stapes in position; skull roof not shown. See Fig. 3c for a reconstruction. **b**, Stereo pair of the ventral view of the braincase and skull roof showing the inflated air chamber, Säve-Söderbergh's process and the right stapes in position. The adductor chamber and marginal teeth can be seen on the morphological

right. See Fig. 3a for a reconstruction. **c**, Stereo pair of the braincase endocast in ventral view with the stapes in position. **d**, Right stapes in medial view. **e**, Right stapes in posterior view. **f**, Interpretive reconstruction of the endocast and stapes in ventral view. Lscc, lateral semicircular canal; V, VII, X, XI, exit foramina for cranial nerves V, VII, X and XI. See Fig. 1 for other definitions.

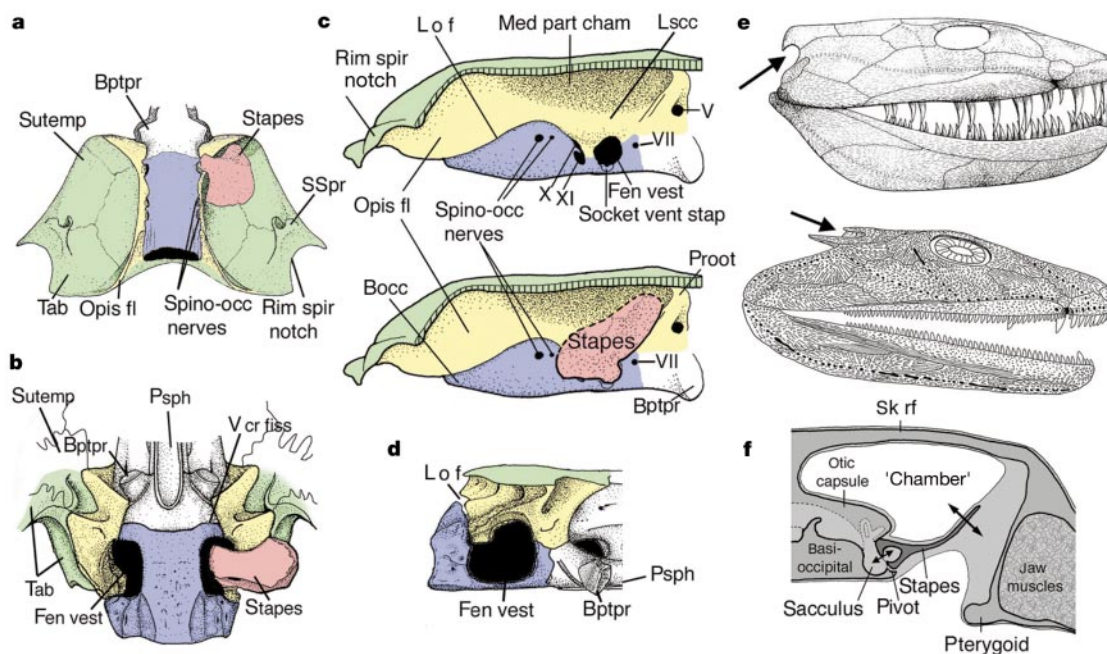


Figure 3 Reconstructions of the braincase and otic region. **a–e**, Comparisons between *Ichthyostega* and *Acanthostega* to show the contrast in structure. **a, b**, Reconstruction of the otic region of the braincase in ventral view, with the pterygoid/epipterygoid removed, to show homologous regions. Note the narrow, elongated otic capsules in *Ichthyostega* (**a**) compared with the short, broad, almost square otic capsules in *Acanthostega* (**b**) and the greatly expanded skull table margin roofing the otic chamber in *Ichthyostega* (for example, compare the areas occupied by the supratemporal and tabular). Furthermore, note the short stout stapes of *Acanthostega* compared with the almost circular stapes of *Ichthyostega*. **c, d**, Reconstructions of the otic regions in lateral view. **c**, Two views of *Ichthyostega* with and without the stapes in place. **d**, *Acanthostega* shown to the same

dorsoventral dimensions to illustrate the difference in length of the otic capsule. Green, dermal skull roof; yellow, otic capsule; blue, basioccipital plus exoccipital; pink, stapes. **e**, Skulls in lateral view to show difference in position of the spiracular notch (arrow). Top, *Ichthyostega*; bottom, *Acanthostega*. **f**, Reconstructed section through the air chamber of *Ichthyostega* to show its relation to the otic capsule and stapes and the possible movement of the stapes. Fen vest, fenestra vestibuli; L o f, lateral otic fissure; Opis fl, opisthotic flange; Rim spir notch, rim of spiracular notch; Socket vent stap, socket for the ventral head of the stapes; Spino-occ nerves, spino-occipital nerves; Sutemp, supratemporal; Tab, tabular bone; V cr fiss, ventral cranial fissure. See Figs 1 and 2 for other definitions.

prevent distortion of the chamber during head movements, breathing and feeding.

We tentatively interpret this ear as an underwater sound transducer, functionally analogous to the swim bladder of ostariophysan fishes (which is connected to the inner ear by so-called weberian

ossicles)¹⁷ or to the middle ear cavity in some aquatic frogs¹⁸. With the head of the animal immersed in water, incoming sound waves would cause the soft ventral wall of the air-filled chamber to vibrate owing to the air–water phase contrast, and the stapes embedded in this wall would transmit the vibrations to the inner ear (Fig. 3f).

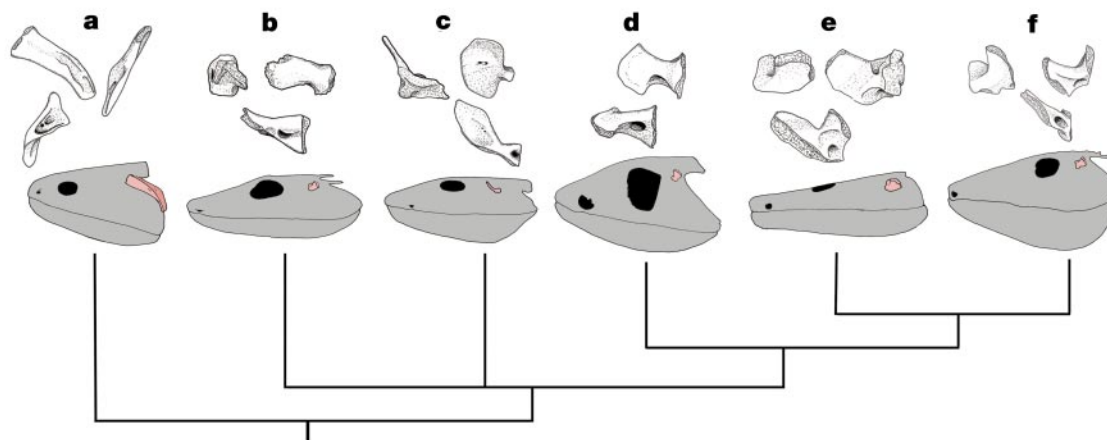


Figure 4 Cladogram of Devonian and Carboniferous taxa showing skulls and hyomandibula/stapes. **a**, *Eusthenopteron* (Late Devonian)^{3,21}; **b**, *Acanthostega* (Late Devonian)^{12,14,16}; **c**, *Ichthyostega* (Late Devonian); **d**, *Pederpes* (Early Carboniferous)⁷; **e**, *Greerpeton* (mid-Carboniferous)^{14,22}; **f**, *Pholiderpeton* (Late Carboniferous)^{13,22}. Skulls (grey) are in left lateral view, with hyomandibula/stapes (pink) in position. The left

hyomandibula/stapes (not to scale) is shown in lateral (top left), dorsal (top right) and posterior (bottom) for each taxon (for *Pederpes*, lateral view is not available in detail). *Acanthostega* and *Ichthyostega* are shown in an unresolved polytomy with the Carboniferous forms.

Although partly speculative, this interpretation best explains the unique morphology of the skull roof, braincase, stapes and stapedial articulation. Furthermore, the recent discovery of deeply grooved branchial arches in *Ichthyostega* (Fig. 1c, d) suggest that, similar to *Acanthostega*¹⁰, it was primarily aquatic rather than the terrestrially adapted creature previously depicted.

Much of the significance of this specialized structure lies in its discovery in what might have been regarded as a primitive tetrapod. This level of specialization so early in the tetrapod fossil record suggests unsuspected diversity among animals that have previously been considered essentially conservative in their cranial anatomy. The ear of *Ichthyostega* is wholly unique, differing markedly from that of its contemporary *Acanthostega*, which seems to represent the ancestral pattern for later tetrapods¹⁶ (Fig. 4). These two are the earliest known tetrapod ears, separated by no more than 10 million years from *Panderichthys*-like ancestors with unmodified fish spiracles and hyomandibulae¹¹. It thus seems that at the origin of tetrapods, ear evolution involved not just a functional shift, but a radical and hitherto unrecognized morphological and functional diversification. The early appearance of a specialized hearing organ in *Ichthyostega* suggests that the otic apparatus had at least some auditory function very early in tetrapod history, on which specializations could be built, although it seems to have related to underwater rather than aerial hearing. □

Methods

High-resolution X-ray CT was used on MGUH f.n. 180. CT-scanned sections were used to create three-dimensional computer models of the objects by reconstructing the surfaces that connected corresponding outlines on adjacent sections^{19,20}.

The otic region and the posterior skull region of the specimen were scanned as two consecutive overlapping data sets with slightly different alignments. We acquired 73 sections of the otic region and 67 sections of the posterior section (on separate occasions). Figure 2a, c–d is based on the anatomically anterior scan set alone, whereas Fig. 2b incorporates both scan sets. Both scans used second-generation (translate-rotate) mode CT, with X-rays set at 420 kV and 1.8 mA and pre-filtered through brass to reduce beam hardening. The parameters of the scans for both series were: 420 kV; 1.8 mA; slice thickness 0.5 mm; interslice spacing 0.5 mm; SOD (source-to-object distance) 752 mm; diameter of field of reconstruction 80 mm (1,024 × 1,024 voxels); voxel size 78 × 78 × 500 μm. Sections were saved as 16-bit TIFF files. Each data set was segmented using Mimics 7.3 software (Materialise N. V.). Because of lateral changes and similarity in the X-ray attenuation values within individual bones and the matrix, local rather than global thresholds were used to create the bone masks (see Supplementary Movies created from these models).

Received 13 May; accepted 11 July 2003; doi:10.1038/nature01904.

1. Säve-Söderbergh, G. Preliminary note on Devonian stegocephalians from East Greenland. *Meddelelser Grönland* **98**, 1–211 (1932).
2. Jarvik, E. On the fish-like tail in the ichthyostegid stegocephalians. *Meddelelser Grönland* **114**, 1–90 (1952).
3. Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1–2 (Academic, New York, 1980).
4. Jarvik, E. The Devonian tetrapod *Ichthyostega*. *Fossils Strata* **40**, 1–206 (1996).
5. Ahlberg, P. E. & Clack, J. A. Lower jaws, lower tetrapods—a review based on the Devonian genus *Acanthostega*. *Trans. R. Soc. Edinb. Earth Sci.* **89**, 11–46 (1998).
6. Laurin, M. The importance of global parsimony and historical bias in understanding tetrapod evolution. Part I. Systematics, middle ear evolution and jaw suspension. *Annales Sci. Naturelles* **1**, 1–42 (1998).
7. Clack, J. A. An early tetrapod from Romer's Gap. *Nature* **418**, 72–76 (2002).
8. Ruta, M., Coates, M. I. & Quicke, D. L. J. Early tetrapod relationships revisited. *Biol. Rev.* **481**, 251–345 (2003).
9. Clack, J. A. & Neining, S. L. in *New Perspectives on the Old Red Sandstone* (eds Friend, P. F. & Williams, B.) 557–566 (Geological Society Symposium Volume, London, 2000).
10. Coates, M. I. & Clack, J. A. Fish-like gills and breathing in the earliest known tetrapod. *Nature* **352**, 234–236 (1991).
11. Ahlberg, P. E., Clack, J. A. & Lukševičs, E. Rapid braincase evolution between *Panderichthys* and the earliest tetrapods. *Nature* **381**, 61–64 (1996).
12. Clack, J. A. The neurocranium of *Acanthostega gunnari* and the evolution of the otic region in tetrapods. *Zool. J. Linn. Soc.* **122**, 61–97 (1998).
13. Clack, J. A. *Pholidrpeton scutigerum* Huxley, an amphibian from the Yorkshire coal measures. *Phil. Trans. Roy. Soc. Lond. B* **318**, 1–107 (1987).
14. Smithson, T. R. The cranial morphology of *Greerpeton burkemorani* (Amphibia: Temnospondylii). *Zool. J. Linn. Soc.* **76**, 29–90 (1982).
15. Bernacke, G. M. & Carroll, R. L. Semicircular canal size in fossil fishes and amphibians. *Can. J. Earth Sci.* **18**, 150–156 (1981).
16. Clack, J. A. Patterns and processes in the early evolution of the tetrapod ear. *J. Neurobiol.* **53**, 251–264 (2002).
17. Yan, H. Y., Fine, M. L., Horn, N. S. & Colon, W. E. Variability in the role of the gasbladder in fish audition. *J. Comp. Physiol.* **186**, 435–445 (2000).

18. Hetherington, T. E. & Lombard, R. E. Biophysics of underwater hearing in anuran amphibians. *J. Exp. Biol.* **98**, 49–66 (1982).
19. Dominguez, P., Jacobson, A. G. & Jefferies, R. P. S. Paired gill slits in a fossil with a calcite skeleton. *Nature* **417**, 841–844 (2002).
20. Ketcham, R. A. & Carlson, W. D. Acquisition, optimization and interpretation of X-ray computed tomographic imagery: Applications to the geosciences. *Comp. Geosci.* **27**, 381–400 (2001).
21. Smithson, T. R. & Thomson, K. S. The hyomandibula of *Eusthenopteron foordi* Whiteaves (Pisces: Crossopterygii) and the early evolution of the tetrapod stapes. *Zool. J. Linn. Soc.* **74**, 93–103 (1982).
22. Clack, J. A. The stapes of the Coal Measures embolomere *Pholidrpeton scutigerum* Huxley (Amphibia: Anthracosauria) and otic evolution in early tetrapods. *Zool. J. Linn. Soc.* **79**, 121–148 (1983).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the staff of MGUH and the Demark and Greenland Geological Survey for cooperation, access to material and support in the field; M. Colbert (UTCT) for technical assistance; H. Blom for discussion of the anatomy of *Ichthyostega*; and NERC and BBSRC for funding the project and additional CT scanning.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.A.C. (j.a.clack@zoo.cam.ac.uk).

Ecological and genetic spatial structuring in the Canadian lynx

Eli Knispel Ruess¹, Nils Chr. Stenseth¹, Mark O'Donoghue², Stan Boutin³, Hans Ellegren⁴ & Kjetill S. Jakobsen¹

¹Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, P.O. Box 1031 Blindern, N-0315 Oslo, Norway

²Fish and Wildlife Branch, Yukon Department of Environment, P.O. Box 310, Mayo, Yukon Y0B 1M0, Canada

³Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada

⁴Department of Evolutionary Biology, Uppsala University, Nordbyvägen 18D, SE 752 36 Uppsala, Sweden

The Canadian lynx, distributed all across the northern part of North America, is well known for its regular population cycles—cycles that have different underlying structures in different parts of Canada¹. Using both nuclear and mitochondrial DNA markers, we report here a close resemblance between the earlier observed spatial ecological structuring of the Canadian lynx¹ and its spatial genetic structuring. Specifically, we demonstrate that the Rocky Mountains represent a barrier to gene flow in western Canada, and, somewhat surprisingly, we detect the presence of a geographically invisible barrier south of Hudson Bay (coinciding with the separation between the ecological Continental and Atlantic regions¹). No evidence for isolation in different glacial refugia within North America was found. We suggest that ecological factors underlying the spatial dynamic structuring also strongly influence the genetic structuring of the Canadian lynx.

The Canadian lynx (*Lynx canadensis*) is distributed fairly continuously throughout the boreal forest of North America. Its main prey species is the equally widely distributed snowshoe hare (*Lepus americanus*). The interaction between snowshoe hares and lynx—itsself affected by hare–vegetation dynamics^{2,3}—is generally believed to generate the well-known regular cycles throughout boreal North America^{1,2–5}. Both species have been the subject of numerous ecological studies (for summaries, see refs 6, 7). The geographical synchrony and ecological structuring have also received much attention^{1,5,8}. Genetic studies (applying microsatellite data) in the

western part of Canada have recently reported high gene flow for both the lynx⁹ and the hare¹⁰. However, no genetic study has thus far covered the entire boreal North American subcontinent, and no study has explored the genetic population structure of lynx within a historical context—a dimension that is important if we are to understand the complex interaction between population genetics and ecology.

We analysed nine microsatellite loci as well as two segments of the mitochondrial genome (cytochrome *b* (*cytb*) and D-loop) for samples of the Canadian lynx, covering its main distribution range. By choosing genetic markers with different rates of evolution, we examined the relative roles played by historical and contemporary influences in shaping the observed spatial structuring. We divided our samples into five large-scale regions: East, Prairie, North, Northwest and British Columbia (BC; Fig. 1), each containing individuals sampled from a number of widespread localities to avoid local effects. Differentiation in microsatellite allele frequencies between regions, measured as pair-wise F_{ST} values, was low but significantly higher than expected under a random distribution of genotypes (see also Supplementary Information), except between the Prairie and North regions (Table 1)—both encompassed by the Continental region of ref. 1. BC was more strongly differentiated than the other regions, regardless of geographical distance (Fig. 2), whereas the remaining point estimates clearly follow a linear increase with geographical distance. This shows that the Rocky Mountains and the adjacent Coastal Mountains represent a barrier to gene flow, not only in the east–west direction, but also to north–south movement. This finding is consistent with studies of radio-collared lynx, showing that long-distance dispersers from the southern Yukon tended to stay north of these mountain ranges and moved northwestward and southeastward rather than south and west into the BC region⁶.

East of the Rocky Mountains there are few, if any, potential large-scale geographical barriers to movement of lynx within the boreal forest of Canada—except for the St Lawrence river (see Supplementary information). However, Stenseth *et al.*¹ reported evidence suggesting that the dynamic structure is divided into three regions

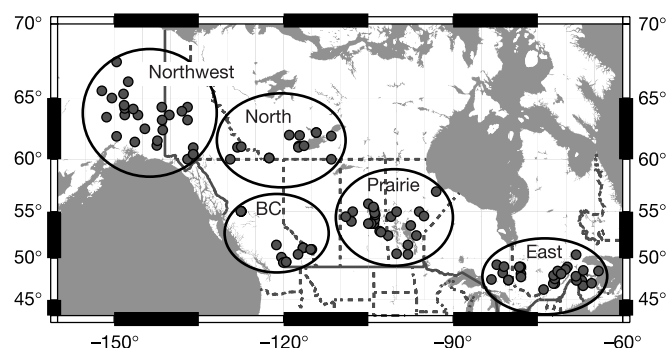


Figure 1 Distribution of the sample localities. Dots, sample localities; dashed lines, province borders; solid line, the Canada–US border. Individual samples (in total 184) were divided on five large-scale geographical regions, as indicated by circles on the map. These were East (Quebec and Ontario) $n = 46$, Prairie (Manitoba and Saskatchewan) $n = 48$, North (NWT and southeast Yukon) $n = 25$, Northwest (Yukon and Alaska) $n = 40$ and BC (British Columbia) $n = 24$ (n is number of individuals). Samples were taken from widespread localities within each region, so these do not represent local populations. Deviations from Hardy–Weinberg equilibrium, heterozygote deficiencies, indicated substructuring within all the regions except BC (see Supplementary Information). The East region corresponds to the Atlantic zone described in ref. 1, the BC region to the Pacific zone, and Prairie and North regions to the Continental zone. Alaska was not covered in ref. 1; it is, however, reasonable to assume that the Northwest region is included in the Continental zone.

Table 1 Genetic differentiation between geographical regions

Microsatellites	Mitochondrial DNA				
	East	Prairie	North	Northwest	BC
East		0.0622**	0.0223*	0.0342*	0.0422*
Prairie	0.0062**		0.0301	−0.0074	0.0119
North	0.0091*	0.0017		0.0261	0.0027
Northwest	0.0156***	0.0095**	0.0027*		0.0073
BC	0.0136**	0.0097**	0.0172**	0.0244***	

Pair-wise F_{ST} estimates for microsatellites below diagonal and mtDNA above diagonal. Significant values indicated as * at the 0.05 level, ** at 0.01 level and *** at the 0.001 level.

across Canada—the Pacific, Continental and Atlantic regions. They argued that this structuring was due to climatic factors—factors that might have profound effects on the winter conditions (such as snow conditions). This ecological structuring may or may not have an effect on the genetic structuring of lynx across Canada. In addition to contemporary ecological processes, genetic population structuring is also a result of various historic processes. During the late Pleistocene, Canada was covered by two extensive ice sheets, separated by the Rocky Mountains, and colonization of the country by biota took place after the retreat of the ice about 12,000 yr before present¹¹. Isolation in different glacial refugia has been shown to have a profound effect on the genetic differentiation within several species^{12,13}. Fossil data, although scarce, indicate that Canadian lynx were present in refugia both in Beringia and south of the ice edge¹⁴. The oldest fossils date back to the Sangamonian interglacial (130–115,000 yr before present) and were found in the southern refugium¹⁴.

Mitochondrial DNA genealogies may allow discrimination between different possible explanations for any detected geographical association¹⁵. The star-like shape of the Canadian lynx mtDNA genealogy (Fig. 3a), with one common haplotype having a central position and recent derivatives independently connected to it by short branches, is just as expected for an abundant species that has expanded its range relatively recently from small or modest numbers of founders¹⁶. Also, the observed mismatch distribution was as

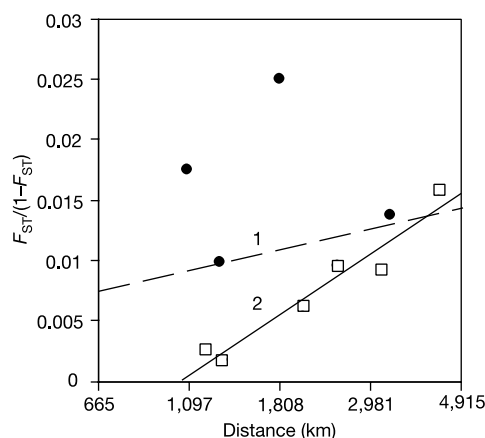


Figure 2 Linearized F_{ST} plotted against geographical distances between samples. Linearized F_{ST} is given by $[F_{ST}/(1 - F_{ST})]$, and is derived from microsatellite data; note logarithmic x-axis. The distances between the centres of the samples ranged between 1,080 km (North and BC) and 4,336 km (East and Northwest). When all samples were compared (curve 1, dashed line), we observed no increase in genetic differentiation with increasing distance between samples (slope, $b = 0.0034$). Removing the BC samples (black dots), however, gave a regression line (curve 2) that is significantly steeper ($b = 0.0098$) than expected under a random distribution of genotypes among regions (Mantel test, $P = 0.044$).

expected after a recent population expansion (raggedness index 0.03, $P = 0.90$), probably corresponding to the post-glacial expansion in North America. None of the major clades was restricted to any geographical region (Fig. 3b), but nested clade analysis¹⁷ suggested restricted gene flow with isolation by distance—indicating that this has taken place after the range expansion, a finding consistent with the above-reported microsatellite results (Fig. 2). The effect of the Rocky Mountains as a barrier to gene flow was not detected in the mtDNA data (Table 1), which might be explained by

the lower rate of evolution in mtDNA than in microsatellites, leading to a slower accumulation of genetic differences in mtDNA. This further suggests that the differentiation we observed with the microsatellite analyses is caused by present rather than historical isolation. Altogether, we found no evidence for past fragmentation in different glacial refugia within North America.

The genetic differentiation between regions in terms of haplotype frequencies, F_{ST} , demonstrates that the eastern region (the Atlantic region of ref. 1) is clearly distinct from all of the other regions (Table 1). The largest F_{ST} value was found between the neighbouring East and Prairie regions, implying that the distinctiveness of the East cannot be explained through an isolation-by-distance effect. As lynx habitat is more or less continuous between the East and the Rocky Mountains, the differentiation cannot be caused by any physical barrier hindering gene flow either. However, Doebeli and Dieckmann¹⁸ have recently demonstrated that genetic diversification may occur along environmental gradients. The differences in climatic conditions between the Atlantic and the Continental region (see ref. 1) may indeed represent such a gradient, leading to the observed genetic structuring.

Although the East region was significantly differentiated from the other regions on the basis of both nuclear and mitochondrial DNA (Table 1), it did not deviate from the general pattern of isolation by distance in the microsatellite data (Fig. 2). This incongruence between patterns indicated by nuclear and mitochondrial markers might be explained by differences in the male and female migration rates, as mtDNA is maternally inherited and hence only reflects the structure of the female population. Stronger differentiation in mtDNA than in nuclear markers would be expected in species with female philopatry. Although male-biased dispersal is characteristic of many species of mammals¹⁹, the data are thus far unclear as to whether this is the case for lynx. Natal philopatry by female kits²⁰ and male-biased immigration²¹ have been noted in some studies, but the most extensive study to date of lynx dispersal²² showed no clear sex bias in dispersal rates or distances.

The level of differentiation in the Canadian lynx is expected to be low owing to the fairly recent population expansion since the latest glaciation, especially given that lynx are periodically abundant and highly mobile. Nevertheless, we have been able to demonstrate the effect of the Rocky Mountains as a geographical barrier in western Canada (a barrier that was not detected in ref. 9), and the existence of a geographically invisible barrier in eastern Canada. Altogether, this may suggest a link between the ecological and genetic spatial structuring across Canada—a link that is probably due to differences in snow conditions¹ and/or differences in the spatio-temporal dynamics of lynx within each of the regions. Future studies aimed at disentangling this link between the ecological and genetic spatio-temporal processes will certainly be rewarding. □

Methods

Microsatellite analysis

The following nine loci from ref. 23 were used: Fca008, Fca031, Fca043, Fca149, Fca391, Fca441, Fca559, Fca628 and F115. Calculations of gene diversity, Hardy–Weinberg tests, F -statistics, the test for population differentiation, regressions for isolation by distance and Mantel tests were done in GENEPop ver. 3.3 (ref. 24).

mtDNA analysis

mtDNA was amplified and sequenced using the following primers. D-loop: mtU, 5'-CTTTGGTCTTGTAACCAAAAAA; and R3, 5'-TAAGAACCAGATGCCAGTA. Cytb: Cytb-1 and Cytb-2 (ref. 25). Altogether, 553 base pairs (bp) of the D-loop and 383 bp of Cytb were sequenced for each individual. Pair-wise F_{ST} estimates were calculated in Arlequin 2000 (ref. 26). The total sample was tested for sudden population expansion by the mismatch distributions approach²⁷ describing the pair-wise differences between sequences, as implemented in Arlequin 2000 (ref. 26). Minimum spanning networks were constructed using the statistical parsimony approach of ref. 17, using the software TCS version 1.13 (ref. 28). Nested clades were identified following the rules of ref. 17, treating

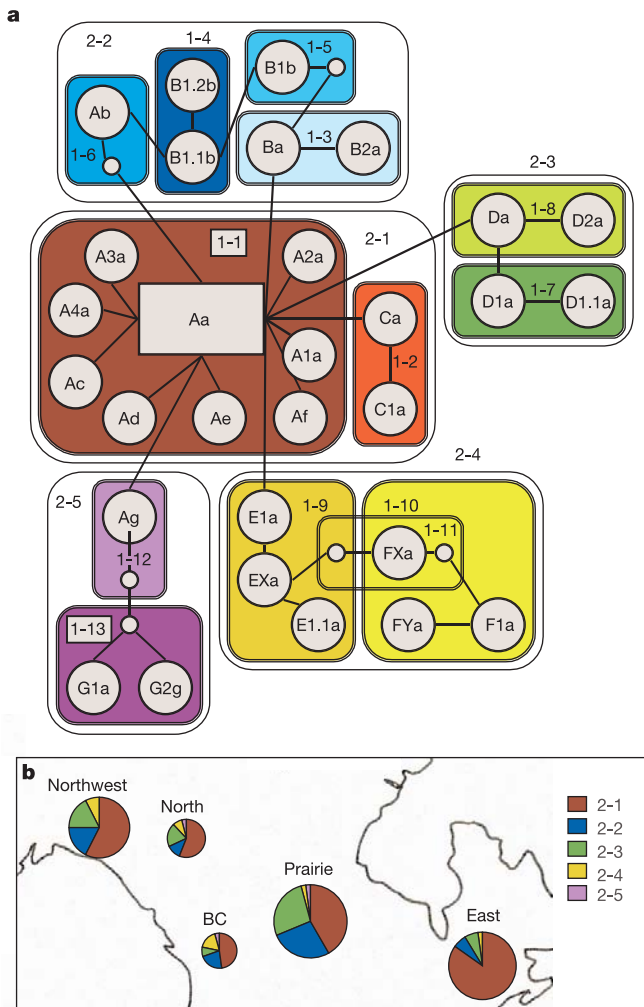


Figure 3 Distribution of mtDNA haplotypes. **a**, Minimum spanning network between mtDNA haplotypes. The D-loop haplotypes are denoted by capital letters and numbers, and Cytb haplotypes are denoted by lower-case letters. Each line represents a single mutational change. Small circles indicate interior nodes that were not present in the sample owing to either insufficient sampling or extinct haplotypes. Similar haplotypes are grouped in clades, shown as colour-coded areas. Loops indicate that more than one equally parsimonious alternative connection exists between haplotypes or clades. Clade 2-1 is the central node of the cladogram, and 59% of the sequenced individuals grouped within this clade; the central haplotype Aa was expressed by 44% of the individuals. **b**, Distribution of clades among geographical regions. The sizes of the pie-charts are proportional to the number of individuals analysed from each region. The network and information about geographical distribution of haplotypes and clades was used for nested clade analysis in order to infer the historical and current processes that probably led to the observed structuring. It is assumed that interior clades are older than tip clades. Restricted gene flow was indicated by the finding that tip clades tend to cover smaller geographical ranges than interior clades (for details, see Supplementary Information).

ambiguities according to ref. 29. The null hypothesis of no geographical association of clades and nested clades was tested by permutation of clades against sampling locations for tip and interior clades in the program GeoDis ver. 2.0 (ref. 30). The biological interpretation of the results was done following the inference key of ref. 15.

Received 21 May; accepted 16 August 2003; doi:10.1038/nature01942.

1. Stenseth, N. C. *et al.* Common dynamic structure of Canada lynx populations within three climatic regions. *Science* **285**, 1071–1073 (1999).
2. Krebs, C. J. *et al.* Impact of food and predation on the snowshoe hare cycle. *Science* **269**, 1112–1115 (1995).
3. Stenseth, N. C., Falck, W., Bjørnstad, O. N. & Krebs, C. J. Population regulation in snowshoe hare and Canadian lynx: Asymmetric food web configurations between hare and lynx. *Proc. Natl Acad. Sci. USA* **94**, 5147–5152 (1997).
4. Elton, C. S. & Nicholson, M. The ten-year cycle in numbers of the lynx in Canada. *J. Anim. Ecol.* **11**, 215–244 (1942).
5. Stenseth, N. C. *et al.* From patterns to processes: Phase and density dependencies in the Canadian lynx cycle. *Proc. Natl Acad. Sci. USA* **95**, 15430–15435 (1998).
6. Mowat, G., Poole, K. G. & O'Donoghue, M. in *Ecology and Conservation of Lynx in the United States* (eds Ruggiero, L. F. *et al.*) 265–306 (University Press of Colorado, Denver, 2000).
7. Krebs, C. J., Boutin, S. & Boonstra, R. *Ecosystem Dynamics of the Boreal Forest: The Kluane Project* (Oxford Univ. Press, New York, 2001).
8. Ranta, E., Kaitala, V. & Lindström, J. Dynamics of Canadian lynx populations in space and time. *Ecography* **20**, 454–460 (1997).
9. Schwartz, M. K. *et al.* DNA reveals high dispersal synchronizing the population dynamics of Canada lynx. *Nature* **415**, 520–522 (2002).
10. Burton, C., Krebs, C. J. & Taylor, E. B. Population genetic structure of the cyclic snowshoe hare (*Lepus americanus*) in southwestern Yukon, Canada. *Mol. Ecol.* **11**, 1689–1701 (2002).
11. Dawson, A. G. *Ice Age Earth: Late Quaternary Geology and Climate* (Routledge, London, 1992).
12. Hewitt, G. M. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc.* **58**, 247–276 (1996).
13. Fedorov, V. B. & Stenseth, N. C. Multiple glacial refugia in the North American Arctic: Inference from phylogeography of the collared lemming (*Dicrostonyx groenlandicus*). *Proc. R. Soc. Lond. B* **269**, 2071–2077 (2002).
14. Kurlén, B. & Anderson, E. *Pleistocene Mammals of North America* (Columbia Univ. Press, New York, 1980).
15. Templeton, A. R., Routman, E. & Phillips, C. A. Separating population structure from population history: A cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–782 (1995).
16. Avise, J. C. *Phylogeography: The History and Formation of Species* (Harvard Univ. Press, Cambridge, Massachusetts, 2000).
17. Templeton, A. R., Crandall, K. A. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619–633 (1992).
18. Doebeli, M. & Dieckmann, U. Speciation along environmental gradients. *Nature* **421**, 259–264 (2003).
19. Greenwood, P. J. Mating systems, philopatry and dispersal in birds and mammals. *Anim. Behav.* **28**, 1140–1162 (1980).
20. Mowat, G. & Slough, B. G. Some observations on the natural history and behavior of the Canada lynx, *Lynx canadensis*. *Can. Field-Nat.* **112**, 32–36 (1998).
21. Slough, B. G. & Mowat, G. Lynx population dynamics in an untrapped refugium. *J. Wildl. Mgmt* **60**, 946–961 (1996).
22. Poole, K. G. Dispersal patterns of lynx in the Northwest Territories. *J. Wildl. Mgmt* **61**, 497–505 (1997).
23. Menotti-Raymond, M. *et al.* A genetic linkage map of microsatellites in the domestic cat (*Felis catus*). *Genomics* **57**, 9–23 (1999).
24. Raymond, M. & Rousset, F. GENEPOP (v 1.2): Population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249 (1995).
25. Janczewski, D. N., Modi, W. S., Stephens, J. C. & O'Brien, S. J. Molecular evolution of mitochondrial 12S RNA and cytochrome b sequences in the pantherine lineage of Felidae. *Mol. Biol. Evol.* **12**, 690–707 (1995).
26. Schneider, S., Roessli, D. & Excoffier, L. *Arlequin Ver. 2.000: A Software For Population Genetic Data Analysis* (Genetics and Biometry Laboratory, Univ. Geneva, Switzerland, 2000).
27. Rogers, A. R. Genetic evidence for a Pleistocene population explosion. *Evolution* **49**, 608–615 (1995).
28. Clement, M., Posada, D. & Crandall, K. A. TCS: A computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1660 (2000).
29. Templeton, A. R. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134**, 659–669 (1993).
30. Posada, D., Crandall, K. A. & Templeton, A. R. GeoDis: A program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.* **9**, 487–488 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to the wildlife managers and trappers that participated in the sampling, especially D. F. McAlpine, R. Lafond, B. Heydon, B. Verbiwski, A. Arsenault, R. Mulders, P. W. Cotter, M. Badry, C. Apps and T. Shury. We also thank Ø. Flagstad and D. Ehrich for discussions. This work was supported by the University of Oslo and the Norwegian Research Council.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to N.C.S. (n.c.stenseth@bio.uio.no). Sequences have been deposited in GenBank under accession numbers AY319482–AY319505 (D-loop) and AY319506–AY319512 (Cytb).

Evolution of cooperation and conflict in experimental bacterial populations

Paul B. Rainey^{1,2} & Katrina Rainey³

¹Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

²School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand

³27 Westwell Road, Auckland, New Zealand

A fundamental problem in biology is the evolutionary transition from single cells to multicellular life forms^{1–3}. During this transition the unit of selection shifts from individual cells to groups of cooperating cells^{1,3,4}. Although there is much theory^{5–15}, there are few empirical studies¹⁶. Here we describe an evolutionary transition that occurs in experimental populations of *Pseudomonas fluorescens* propagated in a spatially heterogeneous environment¹⁷. Cooperating groups are formed by over-production of an adhesive polymer¹⁸, which causes the interests of individuals to align with those of the group. The costs and benefits of cooperation, plus evolutionary susceptibility to defecting genotypes, were analysed to determine conformation to theory^{1,3,12}. Cooperation was costly to individuals, but beneficial to the group. Defecting genotypes evolved in populations founded by the cooperating type and were fitter in the presence of this type than in its absence. In the short term, defectors sabotaged the viability of the group; but these findings nevertheless show that transitions to higher orders of complexity are readily achievable, provide insights into the selective conditions, and facilitate experimental analysis of the evolution of individuality.

Multicellularity has evolved independently on several occasions and is likely to have simple, albeit diverse, explanations^{1,2}. Until now, attention has focused on the advantages of multicellularity and its implications for the development of complexity^{1,2,19–21}. Less consideration has been given to the selective conditions necessary for the evolutionary origin of simple undifferentiated groups: these have special significance because they may have been the raw material for the evolution of multicellular organisms^{2,15,22}.

The origin of cooperating groups of cells requires an understanding of how selection operates at the level of individual cells^{1,3,6,8,12}. Of central importance is the genetic relatedness of the cooperating individuals: if interactions are with relatives then genes causing altruistic or cooperative behaviour can increase in frequency⁵. While costs of cooperation to individual cells are readily envisaged (expression of traits necessary for cohesion, reduced accessibility of clustered cells to nutrients, build-up of toxic metabolic waste) the selective benefit to forming undifferentiated groups of cells is unclear. Size may be an important factor because larger groups of cells are less prone to predation^{2,16,20}; some can migrate further²³. Recent theory suggests that enhanced resource utilization efficiency and reduced interaction with noncooperative individuals are also relevant¹⁵. A related issue concerns the existence of spatial structure¹³, which increases chances for interactions to occur among genetically related cells⁵.

Populations of ancestral smooth (SM) *P. fluorescens* rapidly diversify when propagated in a spatially structured environment (static broth microcosms), generating, via genetic mutation, a range of niche specialist genotypes that are maintained by negative frequency dependent selection¹⁷. One prominent class of niche specialist is the wrinkly spreader (WS), which colonizes the air–liquid interface. Colonization of this niche enables cells to avoid the anoxic conditions that rapidly build up in unshaken broth culture.

Differences in niche preference of ancestral SM and derived WS genotypes (Fig. 1) led to the hypothesis that WS genotypes owe their

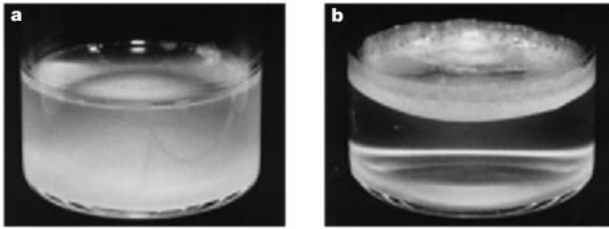


Figure 1 Growth form and niche preference of studied bacteria. **a**, Ancestral (SM) *P. fluorescens* and **b**, derived WS genotypes in static (spatially heterogeneous) microcosms at 25 °C (ref. 17). Only cells of the WS genotype form a mat at the air–broth interface and do so because of a mutation that causes over-production of a polarly expressed polymer. The polymer has glue-like properties and causes daughter cells to remain connected after cell division¹⁸.

evolutionary success to cooperation (Fig. 1b). If true, then according to theory, selection will favour evolution of the cooperating type provided the costs of cooperation to individual cells are traded against increased fitness at the group level^{3–5,11,12}.

To measure fitness of individual WS cells, equal numbers of WS and SM cells were competed over 24 h in spatially heterogeneous microcosms. The ratio of malthusian parameters of each competing type was used to calculate the fitness of WS relative to SM²⁴. During the first 24 h of growth, that is, during the exponential phase (3.5×10^2 to 5.5×10^7 cells ml⁻¹) when resources are abundant, the relative fitness of WS was 0.80 (95% confidence interval, CI: 0.69–0.91, based on the *t* distribution with two degrees of freedom), which is significantly less than 1.0, the fitness by definition of the ancestral genotype.

Despite a much reduced doubling time, WS readily invades (from a single mutant cell) populations dominated by the ancestral genotype to reach population densities that exceed those of the originally dominant ancestral type^{17,25}. Its ability to achieve this is attributable to simple mutations that leads to over-production of a cellulosic polymer¹⁸.

Theory predicts that the group will be vulnerable to the evolutionary emergence of defectors, which, in the absence of conflict mediators, stand to weaken the cohesiveness of the group^{3,4,11,12}. Vulnerability stems from the fact that selection rewards cells that avoid paying the cost of cooperation—selection especially favours defectors that ‘cheat’, that is, cells that gain additional advantage from the cooperating type over and above that gained by avoiding the cost of cooperation.

To test this prediction, replicate microcosms were founded with the WS genotype and their evolutionary trajectory followed over a ten-day period (Fig. 2). By day five, mutants that were ancestral-like in terms of colony morphology and niche preference had evolved by mutation; none of these cells showed any tendency to aggregate or form mats and there was no significant difference in competitive fitness between defectors and ancestral SM cells (mean fitness of defectors relative to ancestral SM was 0.96; 95% CI: 0.86–1.05, based on the *t* distribution with two degrees of freedom). Because these mutants no longer act cooperatively we refer to them as defectors.

We next asked whether defectors were cheats. To do this we examined the population dynamics of WS and defector genotypes during the course of a ten-day period of competition. We then compared these to the dynamics of each genotype propagated on its own (Fig. 3). The fitness of defectors was enhanced (at least during the early stages) when propagated in direct competition with WS. This strongly suggests that defectors, while selectively favoured because of their ability to colonize the broth phase¹⁷, enjoy additional rewards that stem from the presence of WS.

Given that defectors arise by mutation from WS, their origins and fate are intimately associated with the group. A defector that has

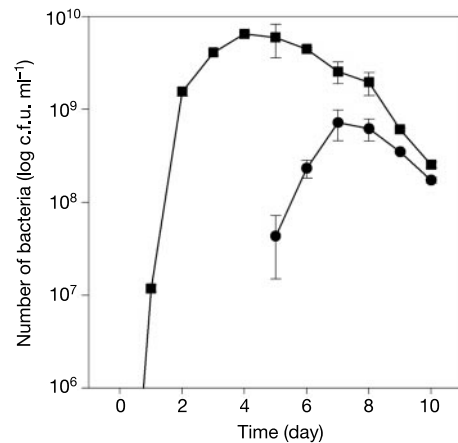


Figure 2 Evolutionary emergence of defecting genotypes from WS during the course of selection in spatially heterogeneous microcosms. The defecting genotypes arise *de novo* by mutation and have a smooth colony morphology that enables them to be readily distinguished from the undulate WS genotypes (see ref. 17). Values are means $\pm$ s.e.m. of three replicates. Squares represent WS; circles represent defecting genotypes. c.f.u., colony-forming units.

avoided the cost of cooperation, but is nevertheless held within the mat is likely to reap a strong frequency-dependent advantage because it can ‘hitchhike’ with WS genotypes, reaping the benefits (especially access to oxygen that is otherwise limiting to defectors), while paying none of the costs. Such behaviour on the part of defectors is evident between days 1 and 2, and days 3 and 4 (Fig. 3a).

In apparent contradiction, defectors appear to be at a selective disadvantage between days 2 and 3, but this is attributable to disproportionate death of defector cells in the broth phase caused by anoxia brought about by mat maturation between days 2 and 3. During this time the WS mat changes from a fine film to a robust structure (Fig. 1b) that restricts oxygen diffusion into the broth phase in a manner akin to a layer of oil. This causes death of a significant fraction of the defector cells in the broth. To show this we harvested bacteria from 2- and 3-day microcosms containing competing WS and defector genotypes and determined the number of cells in the broth phase on each occasion. In 2-day microcosms the broth phase contained on average 6.5×10^9 cells ml⁻¹ (95% CI: 4.0×10^9 to 9.0×10^9 cells ml⁻¹), but dropped to 1.1×10^8 cells ml⁻¹ (95% CI: 2.1×10^8 to 6.5×10^8 cells ml⁻¹) on day 3 (the total number of cells in the mat phase on day 3 being 5×10^9 cells ml⁻¹ (95% CI: 2.2×10^9 to 7.8×10^9 cells ml⁻¹)). Similar reductions in broth-phase colonizers were obtained by applying a 2-mm layer of oil to 2-day cultures. An alternative explanation for the decrease is the activation of a cryptic phage, but extensive searches revealed no evidence of phage activity.

If cheating, via hitchhiking, goes unchecked, the WS group stands to be undermined²⁶. WS genotypes grow unaffected by the presence of cheats through the first four days of selection, even though cheats reap benefit throughout this time (Fig. 3a). In the absence of cheats (Fig. 3b; dotted line) WS gradually declines, primarily due to the success of the mat—it becomes too heavy and gradually sinks into the broth phase where the group suffers the consequences of anoxia. In the presence of cheats, mats collapse abruptly after day 4. This premature collapse is a consequence of the evolutionary success of the cheats and is supported by analyses of the frequency of cheats within the mats: at day 3, 24% of cells within mats are cheats. Because cheats do not contribute to mat integrity their presence is likely to have a negative effect on mat strength. Consistent with this prediction, a cheat-infiltrated mat (24% cheaters) collapses under the weight of 79 mg of glass beads (95% CI: 56–101 mg based on the *t* distribution with two degrees of freedom), whereas a 3-day-old

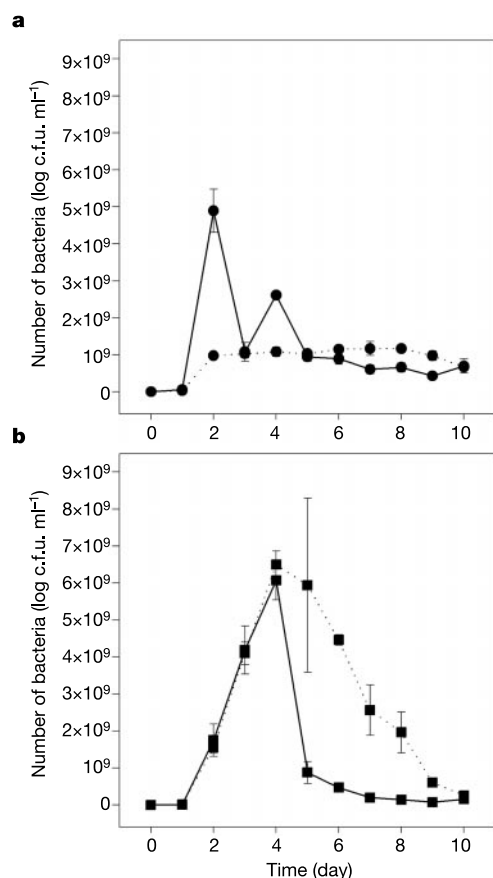


Figure 3 Population dynamics of WS and defector genotypes in the presence and absence of competition. Competing genotypes were founded at equal densities (less than 10^3 cells ml^{-1}). Every 24 h three replicate microcosms were harvested and the frequency of WS and defector genotypes was scored using colony morphology on agar plates to distinguish variant types¹⁷. The effect of WS on the fitness of defector genotypes (**a**) is the difference between dotted (absence of WS) and solid (presence of competing WS) lines. The effect of defector genotypes on the fitness of WS genotypes (**b**) is the difference between the dotted (absence of competition) and solid (presence of competing defector genotype) lines. Values are means $\pm$ s.e.m. ($n = 3$).

mat comprised solely of WS cells supports glass beads to a mass of 432 mg (95% CI: 390–473 mg based on the t distribution with two degrees of freedom).

Here we have described an evolutionary transition from individual cells to a cooperating group that occurs *de novo* during the course of selection of *P. fluorescens* in a heterogeneous environment. The transition is dependent upon spatial heterogeneity; competition for resources (primarily oxygen) is the driving force^{17,25,27}. The cause of cooperation is a cellulosic polymer that is over-produced by WS cells¹⁸. Overproduction of the polymer is costly to individual WS cells, but nevertheless the trait spreads by kin selection⁵ because causing cells (clones) to adhere to one another promotes colonization of the air–liquid interface. Despite the negative impact of defectors on the evolutionary success of the WS mat, WS are a persistent feature of the evolved populations, emerging afresh after each collapse and maintained by negative frequency-dependent selection¹⁷. In all respects, our results confirm crucial elements of long-standing theory^{1,3–6,8,9,11,12,26}.

Undifferentiated groups of WS are a far cry from multicellularity. A likely next step is the evolution of conflict mediators^{1,12,28}. The form of these mediators, and the selective conditions necessary for their emergence, is an experimentally tractable problem; and of some significance because the cellulosic polymer both creates the

group and has the potential to co-evolve with traits that evolve on the basis of the group^{11,29,30}.

Finally, the ease and repeatability of this evolutionary transition is notable. Microbiologists have long known that cultures left on the laboratory bench grow a surface scum analogous to a WS mat. Similar selective forces (competition for oxygen) are likely to operate in a range of aqueous environments, possibly resulting in frequent transitions to undifferentiated multicellularity in the wild. Indeed, an accompanying paper supports this conjecture³¹ and further highlights the significance of adhesive factors²⁹. As such, simple undifferentiated groups of bacteria may have played an important early role in the evolution of multicellularity^{2,15}, but in addition, cooperative behaviour in bacteria may be more common than currently thought. □

Received 5 May; accepted 11 July 2003; doi:10.1038/nature01906.

- Maynard Smith, J. & Szathmari, E. *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
- Bonner, J. T. *First Signals: The Evolution of Multicellular Development*, (Princeton Univ. Press, Princeton, 2000).
- Buss, L. W. *The Evolution of Individuality* (Princeton Univ. Press, Princeton, 1987).
- Maynard Smith, J. in *Evolutionary Progress* (ed. Nitecki, M. H.) 219–230 (Univ. Chicago Press, Chicago, 1988).
- Hamilton, W. D. The genetical evolution of social behaviour. *J. Theor. Biol.* **7**, 1–52 (1964).
- Williams, G. C. *Adaptation and Natural Selection* (Princeton Univ. Press, Princeton, 1966).
- Wilson, D. S. A theory of group selection. *Proc. Natl Acad. Sci. USA* **72**, 143–146 (1975).
- Dawkins, R. *The Selfish Gene* (Oxford Univ. Press, Oxford, 1976).
- Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
- Ferriere, R. & Michod, R. E. The evolution of cooperation in spatially heterogeneous populations. *Am. Nat.* **147**, 692–717 (1996).
- Sober, E. & Wilson, D. S. *Unto Others: The Evolution and Psychology of Unselfish Behaviour* (Harvard Univ. Press, Cambridge, MA, 1998).
- Michod, R. E. *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality* (Princeton Univ. Press, Princeton, 1999).
- Pfeiffer, T., Schuster, S. & Bonhoeffer, S. Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**, 504–507 (2001).
- Smith, J. The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. B* **268**, 61–69 (2001).
- Pfeiffer, T. & Bonhoeffer, S. An evolutionary scenario for the transition to undifferentiated multicellularity. *Proc. Natl Acad. Sci. USA* **100**, 1095–1098 (2003).
- Boraas, M. E., Seale, D. B. & Boxhorn, J. E. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. *Evol. Ecol.* **12**, 153–164 (1998).
- Rainey, P. B. & Travisano, M. Adaptive radiation in a heterogeneous environment. *Nature* **394**, 69–72 (1998).
- Spiers, A. J., Kahn, S. G., Bohannon, J., Travisano, M. & Rainey, P. B. Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. I. Genetic and phenotypic bases of wrinkly spreader fitness. *Genetics* **161**, 33–46 (2002).
- Wolpert, L. The evolution of development. *Biol. J. Linn. Soc.* **39**, 109–124 (1990).
- Bell, G. in *The Origin and Evolution of Sex* (eds Halvorson, H. & Mornoy, A.) 221–256 (Alan R. Liss, New York, 1985).
- Koufopanou, V. & Bell, G. Soma and germ - an experimental approach using *Volvox*. *Proc. R. Soc. Lond. B* **254**, 107–113 (1993).
- Kerszberg, M. & Wolpert, L. The origin of metazoa and the egg: a role for cell death. *J. Theor. Biol.* **193**, 535–537 (1998).
- Foster, K. R., Fortunato, A., Strassmann, J. E. & Queller, D. C. The costs and benefits of being a chimera. *Proc. R. Soc. Lond. B* **269**, 2357–2362 (2002).
- Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341 (1991).
- Buckling, A., Kassen, R., Bell, G. & Rainey, P. B. Disturbance and diversity in experimental microcosms. *Nature* **408**, 961–964 (2000).
- Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).
- Kassen, R., Buckling, A., Bell, G. & Rainey, P. B. Diversity peaks at intermediate productivity in a laboratory microcosm. *Nature* **406**, 508–512 (2000).
- Frank, S. A. Mutual policing and repression of competition in the evolution of cooperative groups. *Nature* **377**, 520–522 (1995).
- Queller, D. C., Ponte, E., Bozzaro, S. & Strassmann, J. E. Single-gene greenbeard effects in the social amoeba *Dictyostelium discoideum*. *Science* **299**, 105–106 (2003).
- Wolf, J. B. Genetic architecture and evolutionary constraint when the environment contains genes. *Proc. Natl Acad. Sci. USA* **100**, 4655–4660 (2003).
- Velicer, G. J. & Yu, Y. N. Evolution of novel cooperative swarming in the bacterium *Myxococcus xanthus*. *Nature* **425** 75–78 (2003).

Acknowledgements We are grateful to D. S. Wilson and A. Buckling whose comments improved the manuscript. A Biotechnology and Biological Sciences Research Council (UK) Research Fellowship to P.B.R. supported the work.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.B.R. (p.rainey@auckland.ac.nz).

Evolution of novel cooperative swarming in the bacterium *Myxococcus xanthus*

Gregory J. Velicer¹ & Yuen-tsu N. Yu²

¹Department of Evolutionary Biology, and ²Department of Protein Evolution, Max-Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany

Cooperation among individuals is necessary for evolutionary transitions to higher levels of biological organization^{1–3}. In such transitions, groups of individuals at one level (such as single cells) cooperate to form selective units at a higher level (such as multicellular organisms). Though the evolution of cooperation is difficult to observe directly in higher eukaryotes, microorganisms do offer such an opportunity⁴. Here we report the evolution of novel cooperative behaviour in experimental lineages of the bacterium *Myxococcus xanthus*. Wild-type strains of *M. xanthus* exhibit socially dependent swarming across soft surfaces⁵ by a mechanism known as ‘S-motility’ that requires the presence of extracellular type IV pili⁶. In lineages of *M. xanthus* unable to make pili, a new mechanistic basis for cooperative swarming evolved. Evolved swarming is mediated, at least in part, by enhanced production of an extracellular fibril matrix that binds cells—and their evolutionary interests—together. Though costly to individuals, fibril production greatly enhanced population expansion in groups of interconnected cells. These results show that fundamental transitions to primitive cooperation can readily occur in bacteria.

Microorganisms display a variety of social behaviours that can enhance fitness in social groups⁴ despite being costly to cooperative individuals in various respects^{7–10}. Simple cooperative groups of undifferentiated cells are likely to have provided the context from which more sophisticated forms of multicellularity first evolved⁸. Direct observations of evolutionary transitions from asocial to cooperative life histories, however, have been lacking. Cooperative S-motility in *M. xanthus* is highly beneficial on some structured surfaces⁵, but is costly to individuals⁷ and easily lost under asocial selective conditions¹¹. Here we ask whether non-cooperative genotypes of *M. xanthus* can evolutionarily regain the ability to swarm effectively on soft surfaces that has been eliminated by a large deletion in a gene essential for wild-type S-motility.

Best known for its construction of fruiting body structures upon starvation¹², *M. xanthus* is a soil-dwelling social predator that uses two genetically distinct motility systems to search for prey¹³. The S-motility system is primarily responsible for group-coordinated movement and requires cell–cell proximity, pili⁶ and an extracellular matrix of cohesive fibrils composed of carbohydrate and proteins¹⁴. Fibrils are also required for normal development¹² and chemotaxis¹⁵. Cell movement in S-motility appears to be driven by pilus extension, adhesion to fibril moieties and subsequent retraction¹⁶. The ‘A-motility’ system allows individualistic cell movement¹³ and appears to be driven by slime extrusion through nozzle-like pores¹⁷. A-motility may also contribute to group movement under some conditions¹⁸. Swarming on soft surfaces is a group function driven primarily by S-motility⁵. A[–]S⁺ (defective A-motility, functional S-motility) mutants swarm proficiently on soft agar, but not on hard agar. A-motility plays a larger role in swarming on harder surfaces, as A⁺S[–] mutants swarm effectively on hard agar but only very poorly on soft agar⁵. The relative roles of A- and S-motility in soil habitats remain unclear.

The capacity for effective group swarming on soft agar via S-motility was eliminated by deleting most of the *pilA* gene encod-

ing pilin, the primary structural component of type IV pili⁶. The resulting A⁺S[–] mutant (‘A1’) and a rifampin-resistant clone with the same deletion (‘A2’) were used to initiate eight independently evolving lineages that were cultured on nutrient soft-agar plates. These ancestors are asocial in the respect that they do not cooperate to swarm effectively on soft agar. At two-week intervals, a small section from the outer edge of each colony swarm was transferred to a fresh plate. After 32 transfer cycles (or 21 in two cases), all eight evolved lines showed significantly greater swarming on soft agar than their respective *pilA* ancestors (Fig. 1a), but two lines (E7 and E8) exhibited particularly dramatic improvements (Fig. 1a, b). The swarm phenotypes of E7 and E8 differ markedly from each other and from the wild-type (‘WT’, A⁺S⁺) parent of their proximate ancestors (A1 and A2, Fig. 1b). Whereas WT radiates evenly outward in a circular swarm, E7 extends outward in a pattern of branching tentacles and E8 expands in broad fans.

The evolved motility phenotypes (hereafter ‘ECM’ for ‘evolved cooperative motility’) are not caused by the evolutionary restoration of pilus production. In both E7 and E8 the ancestral *pilA* deletions are still present, pilin was not detected by anti-PilA antiserum (but was detected in WT), and transmission electron microscopy revealed that pilus structures are present on WT cells

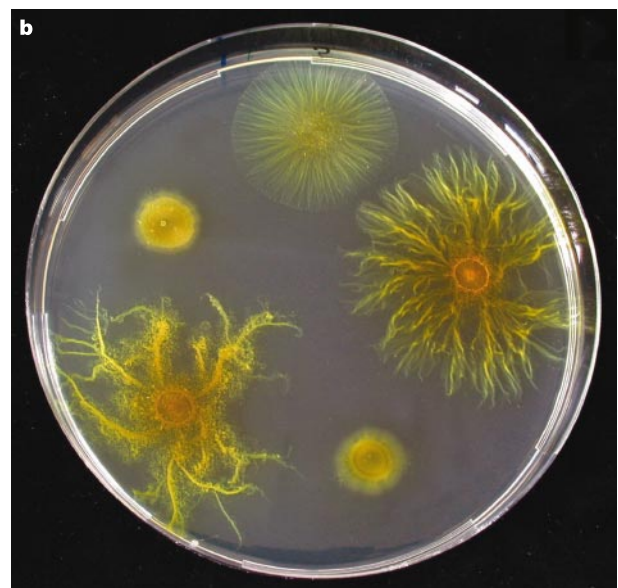
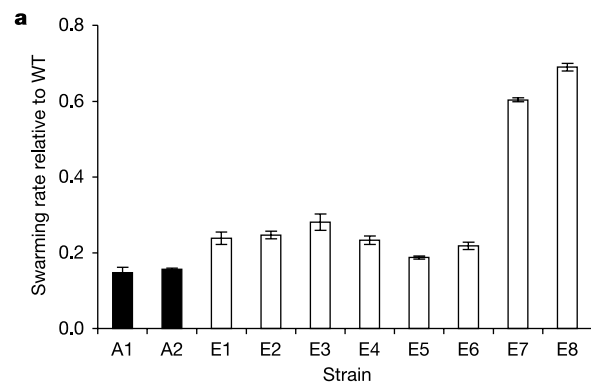


Figure 1 Swarming rates and phenotypes of ancestral (A1, A2) and evolved (E1–E8) genotypes relative to WT on soft agar. **a**, Swarming of evolved populations after 32 cycles of growth (except E5 and E7 at 21 cycles). **b**, Colony swarm phenotypes of WT (top centre), A1, E7, A2 and E8 (anti-clockwise order) after three (WT) or seven (all others) days of swarming on soft agar. E7 and E8 are direct evolutionary descendants of A1 and A2, respectively.

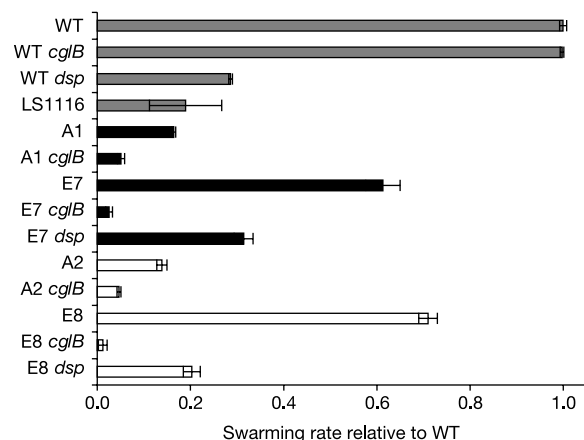


Figure 2 Swarming rates of WT, A1, A2, E7, E8, *cglB* and *dsp* mutants and LS1116 relative to WT on soft agar. Error bars represent 95% confidence intervals.

but are absent from E7 and E8 cells (data not shown). Despite the fact that A-motility cannot drive effective swarming on soft agar in A1 and A2, it is the only known motility system that remains to provide the driving force for ECM in E7 and E8. The necessity of A-motility for ECM is confirmed by the effect of deleting a large segment of the *cglB* A-motility gene¹⁹, which eliminated ECM in both E7 and E8 (Fig. 2). Hypothetically, ECM might have evolved solely by direct modification of the A-motility system, which does not require the presence of fibrils or pili to function. If this were the case, ECM should not be affected by impairment of fibril construction. The point mutation *dsp-1693*²⁰ (hereafter '*dsp*') inhibits fibril production²⁰ and maps to a locus that controls both fibril production and the fibril-mediated detection of chemotactic signals^{21,22}. Introduction of the *dsp* mutation into E7 and E8 greatly reduced ECM (Fig. 2). Moreover, addition of the fibril-inhibiting dye Congo Red²³ to soft agar medium also greatly reduced ECM swarming rates in both E7 and E8, but had very little effect on swarming in E7 *dsp* and E8 *dsp* (data not shown). These results indicate that evolutionary modification of the A-motility system *per se* is not solely responsible for ECM and that fibril production is also involved. The A-motility systems of E7 and E8, however, do appear to have undergone some degree of evolutionary change, because the fibril-deficient *dsp* mutants of these strains show residual swarming that is greater than the swarming of A1 and A2 (Fig. 2).

Because ECM involves fibril production in E7 and E8, defective ECM in the *dsp* mutants of these strains may be subject to extracellular complementation simply due to the presence of their fibril-producing counterparts. Social complementation is in fact observed, as the E7 *dsp* and E8 *dsp* mutants both swarm much further when mixed as a minority with E7 and E8 than they do as isolated genotypes. E7 *dsp* and E8 *dsp* were found to be present at the outer edge of mixed swarms that had migrated 1.9-fold (~17 mm) and 2.8-fold (~29 mm) farther on average, respectively, than clonal swarms of these strains after 11 days. Extracellular complementation of the *dsp*-caused defects in ECM demonstrates the existence of a social component to ECM, irrespective of its mechanistic basis.

Cell-cell cohesion is fundamental to group movement in *M. xanthus*^{5,18}. Production of both pilin and fibrils is required for significant cohesion in WT, as elimination of either prevents normal agglutination into biofilms in static liquid (Fig. 3). While the *pilA* ancestors remain relatively dispersed in static liquid, both E7 and E8 exhibit enhanced cohesion relative to their dispersed-phenotype ancestors (Fig. 3). Because our evolved genotypes lack pilin, their enhanced cohesiveness is likely to have been caused by evolutionary changes in fibril production or composition. Introduction of the *dsp*

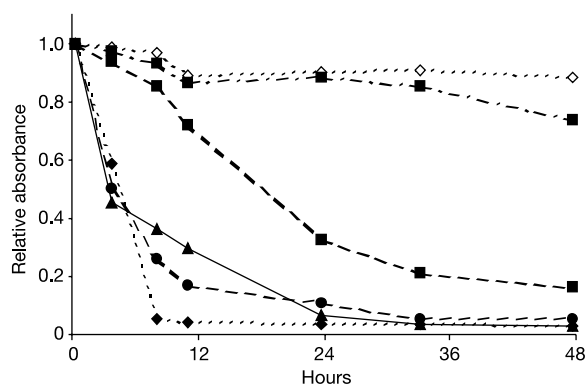


Figure 3 Agglutination patterns of WT (triangles), LS1102 (closed diamonds), WT *dsp* (open diamonds), A1 (squares, long-short dashes), E7 (squares, long-dashes), and E8 (circles).

fibril-deficiency mutation into E7 and E8 eliminates cohesiveness (liquid populations remain dispersed and turbid), the presence of at least two derivatives of the fibril-associated protein FibA²⁴ from whole-cell extracts (Fig. 4a, lanes 6 and 9), and greatly reduces the enhanced fibril production shown by E7 and E8 (Fig. 4b). Moreover, despite their different swarm phenotypes and agglutination patterns (Figs 1b and 3), E7 and E8 share greatly increased levels of 31-kDa and 66-kDa FibA-derived proteins in cell extracts relative to their ancestors (Fig. 4a, lanes 4, 5, 7, 8). Interestingly, the *pilA* deletion in A1 and A2 causes large reductions in fibril material and FibA-derived proteins, thereby indicating that fibril construction is positively regulated by pilin production. The correlation of increased cohesion (Fig. 3) and fibril production (Fig. 4) with ECM, along with the inhibition of these traits both genetically by the *dsp* mutation (Fig. 2) and chemically by the presence of Congo Red, indicates that the social component of ECM involves enhanced fibril production. Fibrils may contribute to ECM by creating an extracellular matrix structure that allows A-motility to drive cells more effectively across soft agar, by mediating the transfer of intercellular chemotactic signals¹⁵, or both. Regardless of its exact mechanism, however, the evolution of ECM represents the evolutionary generation of novel cooperative behaviour in lineages derived from asocial ancestors.

Cooperative behaviours are often costly to the individuals that perform them, and in this system fibril production is costly. Inhibition of fibril production in WT, E7, and E8 by the *dsp* mutation increased the growth rates of these strains in liquid culture by 2.7% (*t*-test, $P = 0.008$), 5.3% ($P = 0.005$), and 4.6% ($P = 0.035$), respectively. Thus, evolved fibril production slows individual reproductive rate under unstructured, asocial conditions, but appears to be favoured by selection at the kin-group level in a structured habitat by its mediation of enhanced group migration. Superior migration allowed clustered groups of related cooperative genotypes to move into territory and consume resources that were unavailable to non-cooperative, poorly swarming competitors.

In principle, a single mutation that enhances fibril expression (and hence cohesion) in an A^+S^- genetic background might allow A-motility to drive significant group swarming on soft surfaces. The A^+S^- strain LS1116¹⁸ is mutationally derived from the parental clone of strain WT and does not make pilin (data not shown), but it carries a mutation (*stk*) that significantly increases fibril expression¹⁸ (Fig. 4a, compare lane 2 to lane 4). This mutation only slightly enhances soft-agar swarming as compared to the improvement shown by E7 and E8 relative to their A^+S^- ancestors (Fig. 2), suggesting that ECM is likely to be caused by a combination of multiple mutations that were incurred during the evolutionary process.

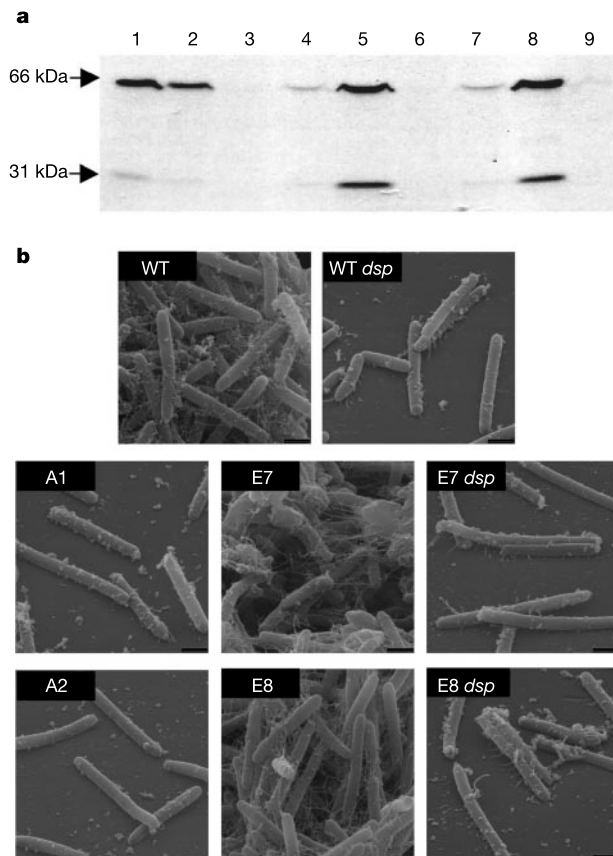


Figure 4 Evolution of fibril production. **a**, Detection of fibril-associated, Fibril-associated, Fibril-associated proteins in whole-cell extracts of WT, LS1116, WT *dsp*, A1, E7, E7 *dsp*, A2, E8, E8 *dsp* (1–9 sequentially). **b**, Images from scanning electron microscopy of cells grown in liquid culture. Large cell clumps bound together by an extensive fibril matrix were common in WT, E7 and E8, but absent in A1, A2 and all *dsp* mutants, cells of which were dispersed. Scale bars represent 1 μ m.

The evolution of cell–cell attachment is required for evolutionary transitions from unicellular to multicellular life histories. Such group cohesion is integral to the social lives of some microorganisms^{25,26}, including pathogenic biofilms²⁷, and to development in multicellular organisms. To be selectively favoured, however, the advantages of any form of group living must outweigh its inherent costs²⁸. In microbes, such advantages may include the enforcement of kinship among cells that share benefits from other cooperative or altruistic traits and the exclusion of non-cooperative genotypes from cooperative groups⁸. In this *Myxococcus* study, production of cohesive fibrils serves to mediate the social component of an evolutionarily new form of cooperative migration. In a conceptually similar transition, evolved cohesion in static liquid cultures of *Pseudomonas fluorescens* allowed colonization of a previously unexploited ecological niche⁹. Together, these studies show that literally ‘sticking together’ can result in quite distinct selective advantages in different organisms. More fundamentally, they directly illustrate the ability of microorganisms to align their evolutionary interests by evolving primitive forms of cooperation. □

Methods

Strains

Strain ‘WT’ in this study is equivalent to strain ‘S’ in refs 7 and 11. Strains LS1102 and LS1116 are described in ref. 18. A1 was made by generating a 561 bp in-frame, markerless deletion (positions 3442–4002 of NCBI sequence L39904) within the *pilA* gene of WT by the method of ref. 29. A2 was constructed in the same manner from a rifampin-resistant clone of WT (strain ‘R’ in refs 7 and 11). Odd- and even-numbered evolution lines derive

directly from A1 and A2, respectively. Whole-population samples from frozen stocks of evolved lineages were used for the swarming assay in Fig. 1a, whereas representative clones of lineages E7 and E8 were used for all other experiments. The *cglB* mutants of Fig. 2 were constructed by the method of ref. 19. *dsp* mutants of WT, E7 and E8 were constructed by transduction of the *dsp-1693* point mutation from strain DK3470¹⁸ into the relevant strain via the *M. xanthus*-specific phage Mx4.

Evolution and swarming assays

Evolution and swarming assays were conducted on 50 ml of CTT (1% casitone, 8 mM MgSO₄, 10 mM Tris-HCl, pH 7.6, 1 mM potassium phosphate, pH 7.6) 0.5% agar medium in 13.6-cm internal-diameter culture plates at 32 °C and 90% relative humidity. Agar medium was allowed to solidify (with plate caps off) in a sterile laminar flow hood and kept overnight (caps on) at room temperature before inoculation. The initial evolution cycle and swarming assays were initiated with 10 μ l culture samples resuspended to 5×10^9 cells ml⁻¹ from centrifuged log-phase cultures grown in CTT liquid at 32 °C and 300 r.p.m. Evolving cultures were transferred after two weeks of swarming by placing an approximately 3 $\times$ 5 mm section from the swarm perimeter upside down on the centre of a fresh plate. Transfer sections were removed from the outermost point on the swarm perimeter for non-circular swarms or from a random perimeter position for circular swarms. This transfer protocol was designed to place positive selection on the ability to reach and use new resources beyond the existing boundaries of a colony swarm. In swarming rate assays, swarm perimeters were marked after one and seven days. Mean swarming rates were calculated for each replicate plate from four evenly distributed vectors (or eight in the case of highly non-circular swarms) at random orientation and for each strain from a minimum of three replicate plates. Swarming rates are shown relative to the WT rate within each separate experiment, which averaged 5.7 mm day⁻¹ across all experiments. The Congo Red swarming experiments were performed by adding Congo Red to CTT soft agar medium at a final concentration of 2.5 μ g ml⁻¹ (for comparison of E7 and E7 *dsp*) or 8 μ g ml⁻¹ (for comparison of E8 and E8 *dsp*). In mixed-genotype swarming assays, E7 *dsp* and E8 *dsp* (both kanamycin-resistant) were mixed at a 1% minority with E7 and E8 (both kanamycin-sensitive), respectively. After 11 days, samples from the swarm edge of five replicate assays (for each strain combination) were placed on agar medium containing kanamycin (40 μ g ml⁻¹) and incubated to test for the presence of the *dsp* genotypes.

Agglutination assay

Liquid cultures ($\sim 5 \times 10^8$ cells ml⁻¹) were removed from shaking (32 °C, 300 r.p.m.) and transferred to static test tubes. Absorbance readings (595 nm) were taken immediately and repeatedly over a 48-h period on a Tecan GENios plate reader. Samples were removed from the uppermost layer of tube cultures, which remained static. Decreasing absorbance represents the settling of agglutinated cell clumps. Absorbance values for each tube were divided by the initial value for that tube. Data points in Fig. 4a are means of three replicates per strain.

Growth rates

Log-phase liquid cultures were diluted to 5×10^4 μ m³ (biovolume) ml⁻¹ ($\sim 5 \times 10^4$ cells ml⁻¹) in 7.5 ml, grown for 24 h, diluted 50-fold into fresh medium, and subsequently transferred (50-fold dilution) daily for five additional days. Biovolume density estimates of each culture were made immediately before dilution transfers using a Multisizer 3 Coulter Counter. The growth rate of each replicate culture was calculated as the rate of ln-transformed biovolume increase, with biovolume estimates adjusted by appropriate dilution factors. Reported growth rate differences were calculated from the means of three independent rate estimates per strain.

Western blot analysis

Whole-cell lysates from $\sim 3.8 \times 10^7$ cells were separated by 10% SDS–PAGE. Proteins were electro-transferred onto PVDF (anti-PilA) or nitrocellulose (Mab2105) and the presence of pilin and fibril-associated proteins were detected by immunoblotting with polyclonal anti-pilA antibody³⁰ and Mab2105²⁴ as primary antibodies, respectively. The immunoblot for pilin was developed with 4-chloronaphthol while the immunoblot for fibril-specific proteins was developed with an ECL luminescence kit (Amersham Pharmacia).

Scanning electron microscopy

Cells were grown to $\sim 5 \times 10^8$ cells ml⁻¹ in CTT liquid, centrifuged, washed in buffer, centrifuged again and fixed with 2.5% glutaraldehyde. Samples were then mounted on polylysine-coated coverslips, postfixed with 1% osmium tetroxide in phosphate-buffered saline, dehydrated in ethanol and critical-point-dried from CO₂. Samples were sputter coated with 8 nm Au/Pd and examined at 20-kV accelerating voltage in a Hitachi S-800 field emission scanning electron microscope.

Received 5 May; accepted 15 July 2003; doi:10.1038/nature01908.

1. Buss, L. W. *The Evolution of Individuality* (Princeton Univ. Press, Princeton, NJ, 1987).
2. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (W.H. Freeman Spektrum, Oxford/New York, 1995).
3. Michod, R. E. & Roze, D. Cooperation and conflict in the evolution of multicellularity. *Heredity* **86**, 1–7 (2001).
4. Crespi, B. J. The evolution of social behavior in microorganisms. *Trends Ecol. Evol.* **16**, 178–183 (2001).
5. Shi, W. & Zusman, D. R. The two motility systems of *Myxococcus xanthus* show different selective advantages on various surfaces. *Proc. Natl Acad. Sci. USA* **90**, 3378–3382 (1993).
6. Wu, S. S. & Kaiser, D. Genetic and functional evidence that Type IV pili are required for social gliding motility in *Myxococcus xanthus*. *Mol. Microbiol.* **18**, 547–558 (1995).
7. Velicer, G. J., Lenski, R. E. & Kroos, L. Rescue of social motility lost during evolution of *Myxococcus xanthus* in an asocial environment. *J. Bacteriol.* **184**, 2719–2727 (2002).

8. Pfeiffer, T. & Bonhoeffer, S. An evolutionary scenario for the transition to undifferentiated multicellularity. *Proc. Natl Acad. Sci. USA* **100**, 1095–1098 (2003).
9. Rainey, P. B. & Rainey, K. Evolution of cooperation and conflict in experimental bacterial populations. *Nature* **425**, 72–74 (2003).
10. Velicer, G. J. Social strife in the microbial world. *Trends Microbiol* **11**, 330–337 (2003).
11. Velicer, G. J., Kroos, L. & Lenski, R. E. Loss of social behaviors by *Myxococcus xanthus* during evolution in an unstructured habitat. *Proc. Natl Acad. Sci. USA* **95**, 12376–12380 (1998).
12. Shimkets, L. J. Intercellular signaling during fruiting-body development of *Myxococcus xanthus*. *Annu. Rev. Microbiol.* **53**, 525–549 (1999).
13. Hodgkin, J. & Kaiser, D. Genetics of gliding motility in *Myxococcus xanthus* (Myxobacterales): two gene systems control movement. *Mol. Gen. Genet.* **171**, 177–191 (1979).
14. Behmlander, R. M. & Dworkin, M. Biochemical and structural analyses of the extracellular matrix fibrils of *Myxococcus xanthus*. *J. Bacteriol.* **176**, 6295–6303 (1994).
15. Kearns, D. B., Campbell, B. D. & Shimkets, L. J. *Myxococcus xanthus* fibril appendages are essential for excitation by a phospholipid attractant. *Proc. Natl Acad. Sci. USA* **97**, 11505–11510 (2000).
16. Li, Y. *et al.* Extracellular polysaccharides mediate pilus retraction during social motility of *Myxococcus xanthus*. *Proc. Natl Acad. Sci. USA* **100**, 5443–5448 (2003).
17. Wolgemuth, C., Hoiczky, E., Kaiser, D. & Oster, G. How myxobacteria glide. *Curr. Biol.* **12**, 369–377 (2002).
18. Dana, J. R. & Shimkets, L. J. Regulation of cohesion-dependent cell interactions in *Myxococcus xanthus*. *J. Bacteriol.* **175**, 3636–3647 (1993).
19. Rodriguez, A. M. & Spormann, A. M. Genetic and molecular analysis of *cglB*, a gene essential for single-cell gliding in *Myxococcus xanthus*. *J. Bacteriol.* **181**, 4381–4390 (1999).
20. Shimkets, L. J. Role of cell cohesion in *Myxococcus xanthus* fruiting body formation. *J. Bacteriol.* **166**, 842–848 (1986).
21. Kearns, D. B. & Shimkets, L. J. Lipid chemotaxis and signal transduction in *Myxococcus xanthus*. *Trends Microbiol.* **9**, 126–129 (2001).
22. Lancero, H. *et al.* Mapping of *Myxococcus xanthus* social motility *dsp* mutations to the *dif* genes. *J. Bacteriol.* **184**, 1462–1465 (2002).
23. Arnold, J. W. & Shimkets, L. J. Inhibition of cell-cell interactions in *Myxococcus xanthus* by Congo Red. *J. Bacteriol.* **170**, 5765–5770 (1988).
24. Kearns, D. B., Bonner, P. J., Smith, D. R. & Shimkets, L. J. An extracellular matrix-associated zinc metalloprotease is required for dilauroyl phosphatidylethanolamine chemotactic excitation in *Myxococcus xanthus*. *J. Bacteriol.* **184**, 1678–1684 (2002).
25. Behmlander, R. M. & Dworkin, M. Extracellular fibrils and contact-mediated cell interactions in *Myxococcus xanthus*. *J. Bacteriol.* **173**, 7810–7821 (1991).
26. Queller, D. C., Ponte, E., Bozzaro, S. & Strassmann, J. E. Single-gene greenbeard effects in the social amoeba *Dictyostelium discoideum*. *Science* **299**, 105–106 (2003).
27. Cramton, S. E., Ulrich, M., Götz, F. & Döring, G. Anaerobic conditions induce expression of polysaccharide intercellular adhesin in *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Infect. Immun.* **69**, 4079–4085 (2001).
28. Krause, J. & Ruxton, G. D. *Living in Groups* (Oxford Univ. Press, Oxford/New York, 2002).
29. Wu, S. S. & Kaiser, D. Markerless deletions of *pil* genes in *Myxococcus xanthus* generated by counterselection with the *Bacillus subtilis* *sacB* gene. *J. Bacteriol.* **178**, 5817–5821 (1996).
30. Wu, S. S. & Kaiser, D. Regulation of expression of the *pilA* gene in *Myxococcus xanthus*. *J. Bacteriol.* **179**, 7748–7758 (1997).

Acknowledgements We are grateful to S. Bonhoeffer, S. Elena, L. Kroos, P. Rainey and W. Shi for discussion or comments, I. Dinkelacker and F. Fiebig for technical assistance, K. Hillebrand for construction of strains A1 and A2, D. Kaiser and H. Kaplan for antibodies and plasmids, L. Shimkets for discussion and strains, and J. Berger and H. Schwarz for electron microscopy expertise and assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.J.V. (gregory.velicer@tuebingen.mpg.de).

Host sanctions and the legume–rhizobium mutualism

E. Toby Kiers¹, Robert A. Rousseau¹, Stuart A. West² & R. Ford Denison¹

¹Agronomy and Range Science, University of California, 1 Shields Avenue, Davis, California 95616, USA

²Institute of Cell, Animal & Population Biology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Explaining mutualistic cooperation between species remains one of the greatest problems for evolutionary biology^{1–4}. Why do symbionts provide costly services to a host, indirectly benefiting competitors sharing the same individual host? Host monitoring of symbiont performance and the imposition of sanctions on ‘cheats’ could stabilize mutualism^{5,6}. Here we show that soybeans

penalize rhizobia that fail to fix N₂ inside their root nodules. We prevented a normally mutualistic rhizobium strain from co-operating (fixing N₂) by replacing air with an N₂-free atmosphere (Ar:O₂). A series of experiments at three spatial scales (whole plants, half root systems and individual nodules) demonstrated that forcing non-cooperation (analogous to cheating) decreased the reproductive success of rhizobia by about 50%. Non-invasive monitoring implicated decreased O₂ supply as a possible mechanism for sanctions against cheating rhizobia. More generally, such sanctions by one or both partners may be important in stabilizing a wide range of mutualistic symbioses.

Mutually beneficial symbiotic relationships between species are ubiquitous, but their evolutionary persistence is puzzling in many cases^{1–3}. If each individual plant or animal host is infected by a single symbiont lineage, then the host and symbiont have a shared interest that may favour cooperation. This is especially so if the symbiont is transmitted vertically, from parent to offspring^{2,7}. However, many mutualisms involve multiple symbiont genotypes per individual host and horizontal transmission of symbionts among unrelated host individuals^{1,2,7}. In this case, each symbiont lineage is selected to increase its own growth and fitness selfishly, at the expense of its host and the other lineages^{2,7}. This is the classic Tragedy of the Commons problem, common to economic and social theory². The tragedy is that while the symbionts as a group could obtain more resources from their host with prudent cooperation, this is not evolutionarily stable because each symbiont lineage gains by selfishly pursuing its own short-term interests.

One possible solution is selection imposed by hosts rewarding

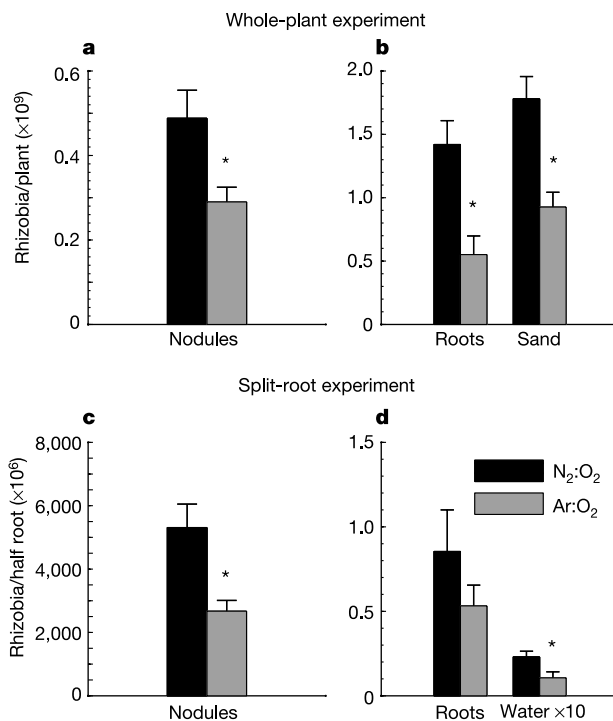


Figure 1 Rhizobia fixing N₂ grew to larger numbers in whole-plant and split-root experiments. Rhizobia allowed (N₂:O₂) or prevented (Ar:O₂) from fixing N₂ by experimental manipulation of atmosphere at the whole-plant (a, b) or split-root (c, d) level were counted (antibiotic media), from nodules (a, c), on the root surface (b, d) and in the surrounding sand (b) or water (d). Counts from water (d) were multiplied by ten for scaling. *Significant differences by ANOVA or paired *t*-test: a, *P* < 0.005, *N* = 11 pairs; b, root fraction, *P* < 0.01, and sand fraction, *P* < 0.01, *N* = 11; c, *P* < 0.001, *N* = 12 plants; d, root fraction, *P* = 0.24, and water fraction, *P* < 0.01, *N* = 12.

cooperation or punishing less cooperative behaviour^{3–6,8,9}. This ‘sanctions’ hypothesis is an evolutionary analogue of the ‘policing’ that can stabilize cooperation within species¹⁰, such as within social insect colonies¹¹. Although sanctions could be important in stabilizing symbiotic mutualisms between species⁸, the difficulty of manipulating most mutualisms experimentally has precluded previous experimental tests of the sanctions hypothesis, or indeed of alternative hypotheses.

Here, we test the sanctions hypothesis with the legume–rhizobium mutualism. Rhizobia are bacteria that fix N_2 within the root nodules of their host legume plants. N_2 fixation is clearly beneficial to the host plant, because it supplies nitrogen needed for growth and photosynthesis. But N_2 fixation (at rates that greatly exceed the nitrogen needs of rhizobia) is energetically costly to the bacteria, and hence reduces the resources that could be allocated to their own growth and reproduction^{5,6}. A single legume plant is typically infected by several different bacterial lineages⁵, creating a potential tragedy of the commons. Consequently, if plants treat fixing and non-fixing nodules similarly (that is, no sanctions), natural selection will favour rhizobia that invest very little in N_2 fixation^{5,6}. Rhizobia vary greatly in the benefits they provide to legumes. Strains that fix little or no N_2 after they form root nodules on legumes are common in some soils^{12,13}. Given the cost of N_2 fixation, why haven’t these cheats completely displaced co-operators? Some mutualisms may be stabilized by the tendency of individuals to associate selectively with better co-operators¹⁴, but legumes cannot consistently recognize and exclude non-fixing rhizobia from infecting their roots^{15,16}.

The legume–rhizobium system offers exceptional opportunities to test the sanctions hypothesis. We can force rhizobia to cheat by replacing air ($N_2:O_2$, 80:20 v/v) with a gas mixture (Ar: O_2 , 80:20 v/v) containing only traces of N_2 (about 0.03% v/v). We estimate that this treatment reduces N_2 fixation to about 1% of normal, based on a K_m (half-saturation N_2 concentration) of about 3%¹⁷. This method allows precise control of when and where rhizobia fix N_2 , without possible confounding effects associated with non-fixing strains. We used this method with soybean (*Glycine max*) and its symbiont *Bradyrhizobium japonicum*. These rhizobia are often mutualistic, but ‘ineffective strains’, which take plant resources but fix little or no N_2 , are widespread^{13,15,16}. We forced rhizobial cheating in: (1) whole plants; (2) one-half of the root system; or (3) individual nodules. In each case, we imposed cheating by exposing target nodules to a nearly N_2 -free atmosphere and exposed control nodules

to air. In the absence of sanctions, we would expect rhizobia fixing little N_2 to direct more resources to their own growth and reproduction. In contrast, if host plants detect the near-cessation of N_2 fixation and apply effective sanctions, then we would predict greater growth and reproduction in the rhizobia allowed to fix N_2 normally.

As predicted by the sanctions hypothesis, forcing rhizobia to cheat by preventing N_2 fixation led to a significant decrease in their fitness. N_2 -fixing rhizobia consistently grew to larger numbers than non-fixing rhizobia in nodules, whether cheating was forced at the plant (Fig. 1a), half-root (Fig. 1c), or nodule level (Fig. 2). In addition, there was a twofold difference (after one plant generation) in release of rhizobia into surrounding sand (Fig. 1b) or nutrient solution (Fig. 1d). Furthermore, rhizobia that had fixed N_2 in nodules had greater survival in sand over five months than rhizobia from the non-fixing treatment (paired t -test, $P < 0.01$, $N = 12$).

The decrease in fitness of the non-fixing rhizobia was associated with a decrease in resource allocation to non-fixing nodules by host plants, as indicated by nodule mass. In experiments where rhizobia were forced to cheat at the half-root or individual-nodule level, each host plant had both fixing and non-fixing nodules, allowing selective partitioning of resources by the host plant. Consistent with the sanctions hypothesis, final nodule fresh weight was higher

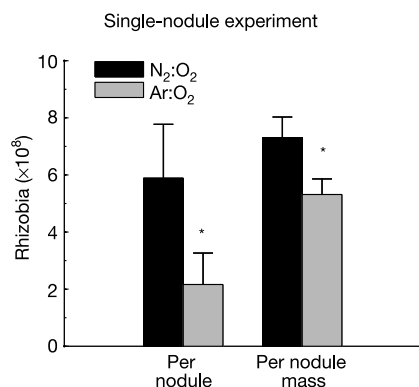


Figure 2 Rhizobia fixing N_2 grew to larger numbers in the single-nodule experiment. Rhizobia allowed to fix ($N_2:O_2$) or prevented from fixing (Ar: O_2) N_2 were counted (antibiotic media) on a per-nodule and per-nodule-mass basis after 10 d of treatment. *Significant differences by paired t -test with $N = 6$ experiments: per nodule, $P < 0.05$; per nodule mass, $P < 0.01$.

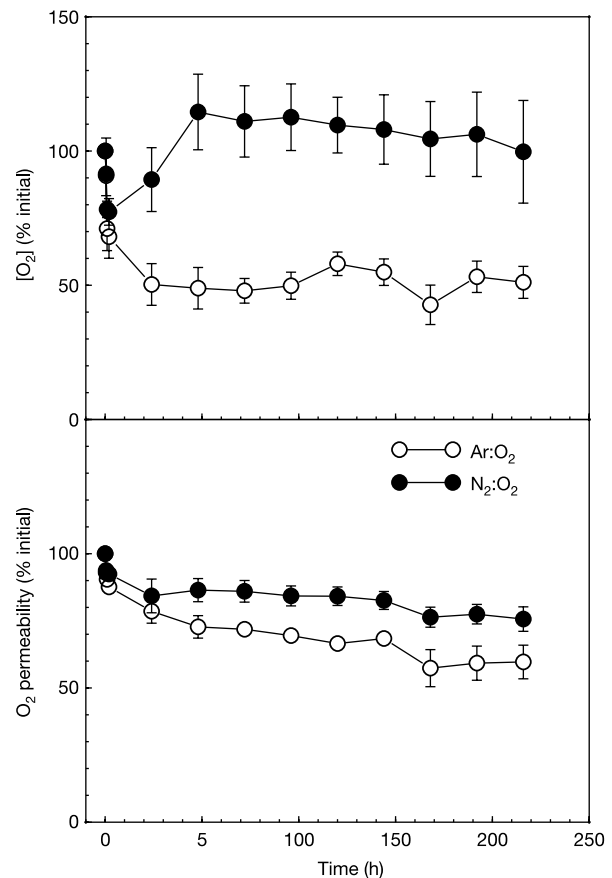


Figure 3 O_2 relations in single nodules where rhizobia were allowed to fix ($N_2:O_2$) or prevented from fixing (Ar: O_2) N_2 . Within 48 h, non-fixing nodules had significantly lower nodule interior O_2 concentration under 20% O_2 , as calculated from leghaemoglobin oxygenation (paired t -test, $P < 0.001$, $N = 6$), and significantly lower O_2 permeability (paired t -test, $P < 0.05$, $N = 6$), relative to controls. Data are presented as % of initial concentration to standardize for any initial differences. A correction for increasing nodule size in controls would have further increased permeability differences between the treatments.

in N_2 -fixing nodules, both in the split-root experiment (paired t -test, $P < 0.001$, $N = 12$) and in the single-nodule experiment (paired t -test, $P < 0.05$, $N = 6$). In addition, root dry weights were higher on the N_2 -fixing side in the split-root experiment (paired t -test, $P < 0.05$, $N = 12$). These results demonstrate how differences in resource allocation^{13,18} at the nodule level are linked to differences in rhizobial fitness.

What is the mechanism by which these sanctions are carried out? Host plants could impose sanctions on non-fixing nodules by attacking rhizobia directly or by decreasing the supply of any resource required for growth^{5,19}. It appears that a decrease in O_2 supply may be the primary mechanism. Nodule interior O_2 concentration and nodule O_2 permeability were both lower in non-fixing nodules within 48 h of the initiation of the experiment (Fig. 3). This decrease in nodule interior O_2 concentration, previously seen in whole-plant experiments²⁰, is the opposite of what would have happened if photosynthate supply had decreased enough to limit respiration in the nodule interior. A lack of significant differences between treatments in O_2 -saturated respiration rate (paired t -test, $P = 0.47$, $N = 6$) also indicated that photosynthate supply did not limit respiration more in non-fixing nodules. Nodule O_2 permeability responds to various conditions that affect nitrogen supply and demand^{21,22}, but responses to soil nitrogen are in the opposite direction (that is, greater O_2 permeability when less nitrogen is available)²³ from the response we found to differences in N_2 fixation. Our results therefore appear to be a specific response to rhizobial defection.

A key assumption in these experiments is that nitrogen supply is unlikely to limit the growth or reproduction of rhizobia in non-fixing nodules directly. Much of the nitrogen needed for nodule growth is imported from the phloem, even in nodules that are exporting much larger quantities of nitrogen to the xylem²⁴. Even under the conservative assumption that plants force complete

nitrogen autonomy on rhizobia in non-fixing nodules, we still estimate that even 1% of the N_2 fixation rate in air would provide enough nitrogen to prevent any direct limitation on rhizobial growth. Specifically, if we assume an N_2 fixation rate in air of 2.6 mg N per g of dry weight of nodule per h (0.276 mmol H_2 and 3 mol H_2 per mol N_2)²⁰ and 2.5 mg bacteroid N per g dry weight of nodule (15 mg bacteroid protein and 6 mg protein per mg N)²⁵, the bacteroids (the differentiated, N_2 -fixing form of rhizobia) in nodules exposed to air would fix enough nitrogen to double in less than an hour. Even at only 1% of the N_2 -fixation rate in air, bacteroids in the Ar: O_2 treatment would fix enough nitrogen to have quadrupled their numbers during the 240 h duration of our single-nodule experiments.

The nodules in our experiment contained only one strain of rhizobia. Mixed nodules can occur, but there is little information on their frequency under field conditions⁵. The potential tragedy of the commons that results from multiple strains per host^{1,3} could also apply to mixed individual nodules. Mixed nodules might reduce the evolutionary effects of nodule-level sanctions if cheats sharing a nodule with mutualists are somewhat protected from nodule-level sanctions⁵. The sanctions reported here are less severe than the flower abortion seen in some yuccas⁹. If rhizobial cheats accumulate more resources than mutualists in the same nodule, as seen by electron microscopy¹⁶, this could perhaps explain the persistence of cheats, despite the fitness cost of cheating in single-strain nodules.

Sanctions directed at specific bacteroids within nodules could be effective in mixed nodules, but only in species in which bacteroids retain the ability to reproduce. Ironically, the most recent evidence for sanctions against bacteroids comes from pea nodules²⁶. In contrast to soybean nodules, bacteroids in pea nodules leave no descendants^{5,27,28}, so denying them resources would have no direct effect on the evolutionary maintenance of cooperation. Only undifferentiated rhizobia, which never fixed N_2 , escape into the soil after pea nodules senesce (Fig. 4). Whole-nodule sanctions, such as cutting off O_2 supply, could affect the survival and reproduction of all rhizobia in the nodule interior. This would impose selection on whichever form is reproductive, and therefore central to the evolution of a given species⁵.

Our results support the hypothesis that legumes select for more cooperative rhizobia by imposing sanctions on the basis of the amount of N_2 that rhizobia fix once established inside nodules. The hypothesis that host sanctions could lead to the evolutionary stabilization of the legume–rhizobium mutualism has been shown previously to be theoretically robust^{6,8}. More generally, sanctions are one way in which the host can control the resource environment of their symbiont, and hence impose a selective environment that favours cooperative behaviour. Mechanisms that can do this, such as sanctions and other more indirect methods^{1,3}, could be important in stabilizing a wide range of mutualistic symbioses. This is because they can favour cooperation when cooperation is otherwise hardest to explain: when there are many symbiont strains per host and there is horizontal symbiont transmission among unrelated host individuals^{1,2,7}. □

Methods

We used an Ar: O_2 atmosphere with only traces of N_2 (about 0.03%, by mass spectrometry) to mimic rhizobial cheats that suddenly stop fixing N_2 . In future experiments, we could alter the timing and composition of gas treatments to simulate rhizobia with different fixation patterns (for example, fixing N_2 at 25% of potential).

Whole-plant experiment

Seeds of a dwarf cultivar of soybean (*Glycine max*; cv. T243, Strain PI 548224, USDA Soybean Germplasm Collection) were sterilized, germinated and planted into autoclaved 700 ml chambers made from stacked Magenta GA-7 culture boxes filled with quartz sand. An air-driven pump recirculated sterile N-free nutrient solution in each chamber. *Bradyrhizobium japonicum* strain USDA 110 ARS was injected into the sand at the base of each seedling, 7 d after planting. Plants were grown with photosynthetically active radiation of $600 \mu\text{E m}^{-2} \text{s}^{-1}$ and 14 h photoperiod. Replicate plants were grouped into four blocks based on acetylene reduction estimates of initial nitrogenase activity and

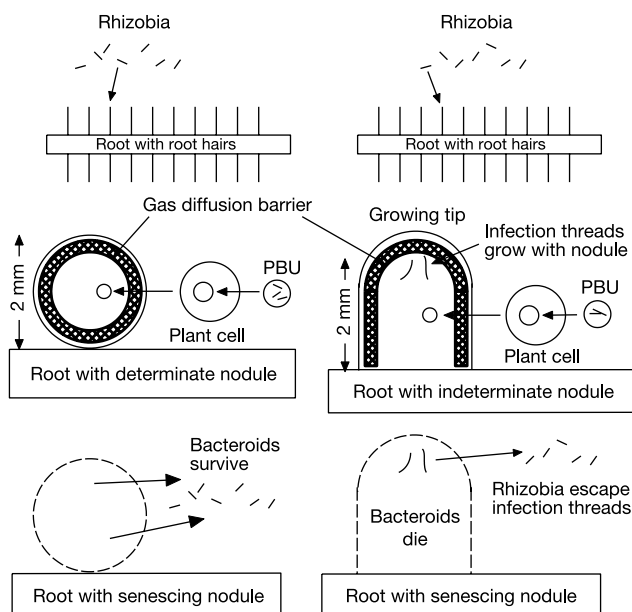


Figure 4 Nodule structure and the life history of rhizobia. After rhizobia differentiate into N_2 -fixing bacteroids, they lose the ability to reproduce in nodules with indeterminate growth (for example, pea), but not in nodules with determinate growth (for example, soybean)⁵. The gas permeability of the O_2 diffusion barrier is under the control of the plant in both types of nodule^{21,22}. The peribacteroid unit (PBU) consists of one or more bacteroids surrounded by a plant membrane, which could perhaps allow sanctions at the bacteroid level.

randomly assigned to N₂-fixing and non-fixing treatments. Either N₂:O₂ or Ar:O₂ was delivered through perforated plastic tubing 1 cm above the base, at 100 ml min⁻¹. Three months after planting, nodules were removed from roots. Roots were cut, vortexed, and sonicated in a FS20 'watch-bath' type sonicator in 0.01% Tween 20. The extractant was diluted 10⁶-fold and spread on MAG antibiotic-containing plates. Intact nodules were removed from roots, counted, weighed and crushed in a tissue homogenizer, diluted and plated. Sand from each box was homogenized for 30 min in a sterile flask containing sterile 0.01% Tween 20, on a flask rotator. A liquid subsample was removed from the sand mixture 3 cm below the water line, diluted by 10⁴ and plated. Colonies grew for 10 d at 32 °C and colony-forming units (c.f.u., mean of eight plates) were recorded.

Split-root experiment

Seeds of *G. max* semidwarf variety 'S0066' were sterilized, germinated, and inoculated with approximately 10⁷ cells per seedling. Twelve plants, each with two similar root halves (resulting from regrowth after root-tip removal), were transplanted to hydroponic chambers, with similar nodule numbers on each half of a chamber divided by a silicone gel seal. Chamber halves were randomized into two treatments, either N₂:O₂ or Ar:O₂ (80:20, v/v) at 130 ml min⁻¹, 5 d after transplanting. H₂ production was measured to confirm disruption of N₂ fixation²⁹ by Ar:O₂. Five weeks after transplanting, roots, nodules and rhizobia in nutrient solution were processed as described above for the 12 replicates, each a paired comparison. For survival assays, nodule homogenate was diluted and added at an estimated 10⁵ rhizobia per g sterile sand. Twenty weeks later, rhizobial populations were determined by plate counts.

Single-nodule experiment

Six independent replicate experiments used *G. max* 'S0066' grown in plastic growth pouches and inoculated as above. Fifteen days later, two nodules of equal size were selected per plant. Fixing and non-fixing treatments were randomized. Chambers of 2 cm diameter were positioned around intact nodules, with 250 ml min⁻¹ of humidified N₂:O₂ or Ar:O₂ flowing through each chamber. Fractional oxygenation of leghaemoglobin under air, nodule O₂ permeability, and O₂-saturated respiration rate were measured daily as previously described^{23,30}. Briefly, nodules were exposed successively to 20, 0, 70 and 0% O₂ while fractional oxygenation of the nodule protein leghaemoglobin was measured by non-invasive spectrophotometry. O₂ permeability was calculated from the rate of increase in oxygenation after switching to 70% O₂, after correcting for respiration, which was calculated from the rate of oxygenation decrease as interior O₂ fell from O₂-saturated to O₂-limited concentrations after switching to 0% O₂. After 10 d, nodules were weighed, crushed, and assayed for c.f.u. per nodule and per g of nodule. Analyses of variance and Tukey's studentized range test for whole-root, and paired *t*-tests for split-root and single nodule experiments were conducted using SAS software (SAS Institute).

Received 9 May; accepted 18 July 2003; doi:10.1038/nature01931.

- Herre, E. A., Knowlton, N., Mueller, U. G. & Rehner, S. A. The evolution of mutualisms: exploring the paths between conflict and cooperation. *Trends Ecol. Evol.* **14**, 49–53 (1999).
- Frank, S. A. *Foundations of Social Evolution* (Princeton Univ. Press, Princeton, 1998).
- Yu, D. W. Parasites of mutualisms. *Biol. J. Linn. Soc.* **72**, 529–546 (2001).
- Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
- Denison, R. F. Legume sanctions and the evolution of symbiotic cooperation by rhizobia. *Am. Nat.* **156**, 567–576 (2000).
- West, S. A., Kiers, E. T., Simms, E. L. & Denison, R. F. Sanctions and mutualism stability: why do rhizobia fix nitrogen? *Proc. R. Soc. Lond. B* **269**, 685–694 (2002).
- Crespi, B. J. The evolution of social behavior in microorganisms. *Trends Ecol. Evol.* **16**, 178–183 (2001).
- West, S. A., Kiers, E. T., Pen, I. & Denison, R. F. Sanctions and mutualism stability: when should less beneficial mutualists be tolerated? *J. Evol. Biol.* **15**, 830–837 (2002).
- Pellmyr, O. & Huth, C. J. Evolutionary stability of mutualism between yuccas and yucca moths. *Nature* **372**, 257–260 (1994).
- Frank, S. A. Mutual policing and repression of competition in the evolution of cooperative groups. *Nature* **377**, 520–522 (1995).
- Ratnieks, F. L. W., Monnin, T. & Foster, K. R. Inclusive fitness theory: novel predictions and tests in social Hymenoptera. *Ann. Zool. Fennici* **38**, 201–214 (2001).
- Burdon, J. J., Gibson, A. H., Searle, S. D., Woods, M. J. & Brockwell, J. Variation in the effectiveness of symbiotic associations between native rhizobia and temperate Australian *Acacia*: within-species interactions. *J. Appl. Ecol.* **36**, 398–408 (1999).
- Singleton, P. W. & Stockinger, K. R. Compensation against ineffective nodulation in soybean. *Crop Sci.* **23**, 69–72 (1983).
- Ferriere, R., Bronstein, J. L., Rinaldi, S., Law, R. & Gauduchon, M. Cheating and the evolutionary stability of mutualisms. *Proc. R. Soc. Lond. B* **269**, 773–780 (2001).
- Amarger, N. Competition for nodule formation between effective and ineffective strains of *Rhizobium meliloti*. *Soil Biol. Biochem.* **13**, 475–480 (1981).
- Hahn, M. & Studer, D. Competitiveness of a *nif⁻* *Bradyrhizobium japonicum* mutant against the wild-type strain. *FEMS Microbiol. Lett.* **33**, 143–148 (1986).
- Rasche, M. E. & Arp, D. J. Hydrogen inhibition of nitrogen reduction by nitrogenase in isolated soybean nodule bacteroids. *Plant Physiol.* **91**, 663–668 (1989).
- Singleton, P. W. & van Kessel, C. Effect of localized nitrogen availability to soybean half-root systems on photosynthate partitioning to roots and nodules. *Plant Physiol.* **83**, 552–556 (1987).
- Udvardi, M. K. & Kahn, M. L. Evolution of the (*Brady*)*Rhizobium*-legume symbiosis: why do bacteroids fix nitrogen? *Symbiosis* **14**, 87–101 (1993).
- King, B. J. & Layzell, D. B. Effect of increases in oxygen concentration during the argon-induced decline in nitrogenase activity in root nodules of soybean. *Plant Physiol.* **96**, 376–381 (1991).
- Sheehy, J. E., Minchin, F. R. & Witty, J. F. Biological control of the resistance to oxygen flux in nodules. *Ann. Bot.* **52**, 565–571 (1983).

- Hartwig, U., Boller, B. & Nösberger, J. Oxygen supply limits nitrogenase activity of clover nodules after defoliation. *Ann. Bot.* **59**, 285–291 (1987).
- Denison, R. F. & Harter, B. L. Nitrate effects on nodule oxygen permeability and leghemoglobin. Nodule oximetry and computer modeling. *Plant Physiol.* **107**, 1355–1364 (1995).
- Layzell, D. B., Rainbird, R. M., Atkins, C. A. & Pate, J. S. Economy of photosynthetic use in nitrogen-fixing legume nodules. *Plant Physiol.* **64**, 888–891 (1979).
- Sen, D. & Weaver, R. W. Nitrogen fixing activity of rhizobial strain 32H1 in peanut and cowpea nodules. *Plant Sci. Lett.* **18**, 315–318 (1980).
- Lodwig, E. M. *et al.* Amino-acid cycling drives nitrogen fixation in the legume–*Rhizobium* symbiosis. *Nature* **422**, 722–726 (2003).
- Kijne, J. W. The fine structure of pea root nodules. 2. Senescence and disintegration of the bacteroid tissue. *Physiol. Plant Pathol.* **7**, 17–21 (1975).
- Sprent, J. I. & Raven, J. A. Evolution of nitrogen-fixing symbioses. *Proc. R. Soc. Edinb. B* **85**, 215–237 (1985).
- Layzell, D. B., Hunt, S., King, B. J., Walsh, K. B. & Weagle, G. E. in *Applications of Continuous and Steady-State Methods to Root Biology* (eds Torrey, J. G. & Winship, L. J.) 1–28 (Kluwer Academic, Dordrecht, 1989).
- Denison, R. F. & Layzell, D. B. Measurement of legume nodule respiration and O₂ permeability by noninvasive spectrophotometry of leghemoglobin. *Plant Physiol.* **96**, 137–143 (1991).

Acknowledgements We thank R. Grosberg, S. Nee, A. Griffin, D. Shuker and M. Stanton for comments, P. Graham, D. Phillips and M. King for advice on methods, R. Nelson and D. Smith for soybean cultivars 'T243' and 'S0066', and P. van Berkum for *B. japonicum* 110ARS. This work was supported by the NSF (grant to R.E.D. and graduate fellowship to E.T.K.), the California Agricultural Experiment Station, the Land Institute, the Royal Society, the BBSRC and the NERC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.E.D. (rfdenison@ucdavis.edu).

Functional genetic analysis of mouse chromosome 11

Benjamin T. Kile^{1*}, Kathryn E. Hentges^{1*}, Amander T. Clark^{1*}, Hisashi Nakamura¹, Andrew P. Salinger¹, Bin Liu¹, Neil Box¹, David W. Stockton¹, Randy L. Johnson², Richard R. Behringer³, Allan Bradley¹ & Monica J. Justice¹

¹Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

²Department of Biochemistry and Molecular Biology, and ³Department of Molecular Genetics, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA

*These authors contributed equally to this work

† Present address: The Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

Now that the mouse and human genome sequences are complete, biologists need systematic approaches to determine the function of each gene^{1,2}. A powerful way to discover gene function is to determine the consequence of mutations in living organisms. Large-scale production of mouse mutations with the point mutagen N-ethyl-N-nitrosourea (ENU) is a key strategy for analysing the human genome because mouse mutants will reveal functions unique to mammals, and many may model human diseases³. To examine genes conserved between human and mouse, we performed a recessive ENU mutagenesis screen that uses a balancer chromosome, inversion chromosome 11 (refs 4, 5). Initially identified in the fruitfly, balancer chromosomes are valuable genetic tools that allow the easy isolation of mutations on selected chromosomes⁶. Here we show the isolation of 230 new recessive mouse mutations, 88 of which are on chromosome 11. This genetic strategy efficiently generates and maps mutations on a single chromosome, even as mutations throughout the genome are discovered. The mutations reveal new defects in haematopoiesis, craniofacial and cardiovascular development, and fertility.

Table 1 Eighty-eight new mutants on mouse chromosome 11

Category	Number new*	Locus name	Phenotype	Number known†
Lethal	6	<i>I11Jus1,-2,-3,-4,-7,-12</i>	Prenatal death post implantation, before E9.5	5
	12	<i>I11Jus5,-6,-8,-9,-11,-14,-17-20,-27,-39</i>	Prenatal death E9.5–12.5	6
	2	<i>I11Jus15,-29</i>	Prenatal death after E12.5	3
	10	<i>I11Jus24,-30,-32,-34-38,-42,-46</i>	Prenatal death, stage not determined	
	23	<i>I11Jus10,-13,-16,-21-23,-25,-26,-31,-33,-40,-41,-43,-44,-47,-48,-51-57</i>	Postnatal death	8
	2	<i>I11Jus28,-45</i>	Time of death unknown	
Neurological	10	<i>nur1,-2</i>	Nervousness, tremors	6
		<i>nur3,-6</i>	Small, hyperactive	
		<i>nur4,-5</i>	Tremors	
		<i>nur7</i>	Small, lethargic, tremors in adults	
		<i>nur8</i>	Epilepsy, craniofacial, eye	
		<i>nur9</i>	Weaving head/hearing loss	
		<i>nur10</i>	Small, splayed gait	
		<i>crt2,-3,-5,-8,-12</i>	Craniofacial abnormalities	1‡
		<i>skc1,-2</i>	Coat texture	1§
		<i>gro1</i>	Small at P9 to adult	2
Skeletal: craniofacial	5	<i>gro22</i>	Small, low cholesterol	
Skin/coat	2	<i>gro40</i>	Small, half size	
Growth/size	5	<i>gro41</i>	Small, sickly	
		<i>gro42</i>	Small, two-thirds size	
Infertility	6	<i>inf2,-4-6,-8,-9</i>	Male infertility	1
	2	<i>inf3,-7</i>	Female infertility	0
Haematopoietic	3	<i>hem1,-3</i>	Immature WBCs	2
		<i>hem2</i>	Anaemia	

Locus names are derived from the underlined letters.
*The number of each phenotype isolated in this ENU screen. These mouse mutants are available to the scientific community on request at <http://www.mouse-genome.bcm.tmc.edu>.
†The number of previously reported spontaneous or targeted mutants that lie in the *Trp53–Wnt3* region²⁸. An additional six mutations had phenotypes detected by assays not performed in this screen to make a total of 45 mutations with phenotypes reported. Twenty-one targeted mutations in this region were reported as having no phenotype.
‡Four additional mutations with skeletal phenotypes were reported.
§One recessive coat colour mutation was reported.

Genetic strategies that can isolate mouse mutations in regions that are conserved between the mouse and human provide ways to infer the function of human genes. Human chromosome 17 genes that are conserved in the mouse are located on mouse chromosome 11. The inversion chromosome 11 (Inv(11)8Brd^{*Trp53–Wnt3*}; Fig. 1a)

balancer has properties that allow us to isolate mutations on chromosome 11 after a three-generation breeding scheme (Fig. 1b), even though mutations induced by ENU will occur throughout the genome of the treated male (genome-wide mutations recovered are summarized in Supplementary Table 1).

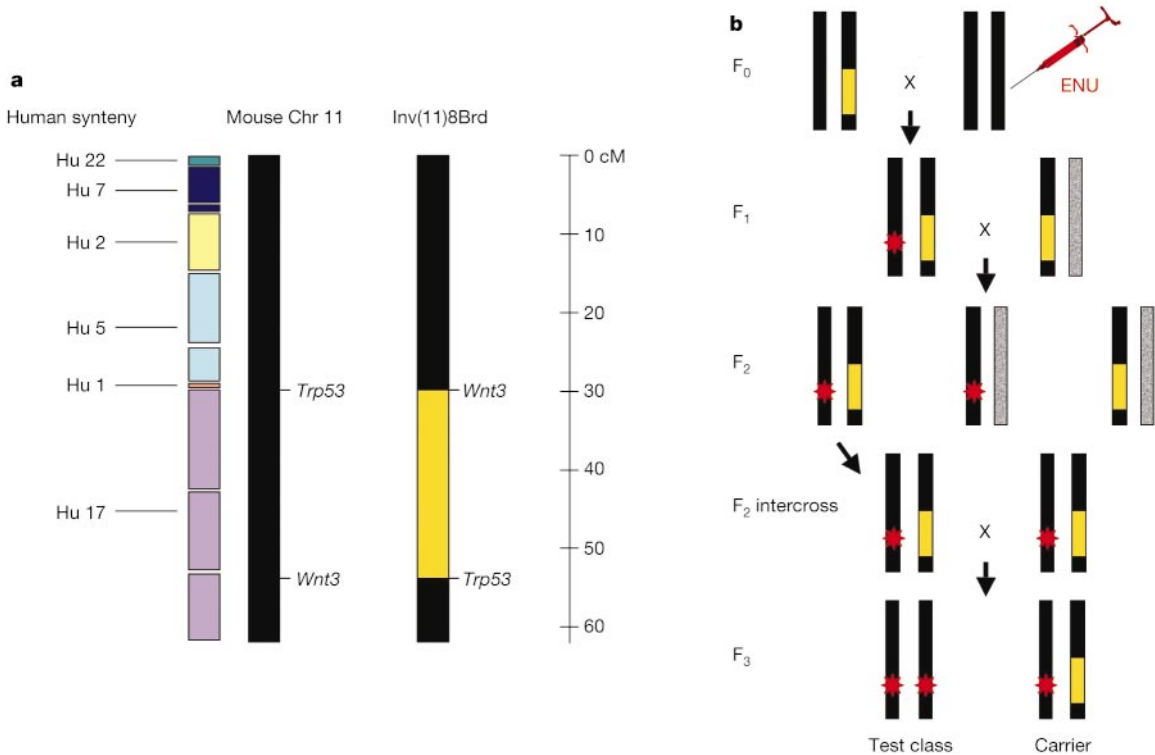


Figure 1 A balancer chromosome screen for the conserved region of mouse chromosome 11/human chromosome 17. **a**, The distal two-thirds of mouse chromosome 11 is conserved with human chromosome 17 (purple)¹. The Inv(11)8Brd^{*Trp53–Wnt3*} balancer contains a 24-cM inversion (34 Mb; yellow box) between *Trp53* and *Wnt3* on mouse chromosome 11. **b**, ENU-treated C57BL/6J males (black chromosomes) are mated to females carrying the balancer (yellow box; *K14-Agouti* produces yellow ears, tail and

ventrum) to generate first generation (F₁) animals that may carry a new mutation (red asterisk). F₁ animals are mated to mice carrying a balancer and *Rex* (dominant curly coat, grey chromosome), which marks the non-mutagenized chromosome. Informative F₂ animals identified by their yellowish colour and lack of curly coat are mated to look for new mutations (two asterisks; test class) in the F₃ offspring. Mice homozygous for the inversion are *Wnt3*-deficient and die by E10.5 (not shown).

For example, *Inv(11)8Brd^{Trp53-Wnt3}* contains a 24-cM inversion on mouse chromosome 11 between the *Trp53* and *Wnt3* genes that suppresses the recovery of recombination products over nearly half of the chromosome. Thus, mutations induced in a wild-type chromosome will remain fixed when maintained *in trans* to the balancer in heterozygous animals. The inversion has a dominant *Agouti* coat colour marker, which allows animals that carry it to be easily identified. Therefore, heterozygotes and mutant homozygotes can be easily distinguished by their coat colour.

In contrast to previous mouse recessive ENU screens that focused on a single stage of embryogenesis^{7–9}, our genetic screen using the *Inv(11)8Brd^{Trp53-Wnt3}* balancer is designed to isolate all lethal mutations from fertilization to adulthood, because the simple absence of test class animals (those lacking the *Agouti* transgene) identifies a lethal mutation. Fifty-five of the mouse chromosome 11 mutants are lethal when homozygous (Table 1). Thirty pedigrees carry mutations that cause prenatal death, whereas 23 pedigrees carry mutations that cause postnatal death (Supplementary Table 2). Many of the postnatal lethal mutations exhibit visible pheno-

types such as small size, developmental delay, craniofacial or neurological abnormalities (Supplementary Fig. 1 and movie). Lethal mutations are the most common phenotypic class isolated in model organisms such as *Caenorhabditis elegans* and *Drosophila*^{10,11}. Approximately 60% of the mouse chromosome 11 mutations identified in this screen are lethal in homozygotes, and nearly one-half of the spontaneous or targeted mutations in the balancer interval are lethal (Table 1). Two mutagenesis screens that would isolate mutations in a defined genomic region have been carried out previously in the mouse, and each of these screens yielded primarily lethal mutations^{12,13}. Together, these data suggest that many mouse mutations are detrimental to the survival of the animal. One end of *Inv(11)8Brd^{Trp53-Wnt3}* disrupts *Wnt3*, causing embryonic lethality in mice homozygous for the balancer. Combined with the property of recombination suppression, this allows lethal mutations to be easily maintained *in trans* through heterozygous crosses, without requiring molecular markers for stock maintenance.

Further analysis of the cause of death of the lethal mutations revealed a broad range of distinctive phenotypes at every stage of development (Fig. 2; see also Supplementary Fig. 2). Some chromosome 11 mutants exhibited abnormal somite morphology or neural tube defects (*l11Jus5*, -8, -15, -39), which are often observed in human birth defects. Others, such as *l11Jus3*, had abnormal visceral endoderm cells and failed to gastrulate (Fig. 2a, b). However, most of the phenotypes are the result of defects in organs and tissues such as heart, blood vessels, placenta, or haematopoietic cells. For example, *l11Jus27* mutants have distended tubular hearts (Fig. 2c, d). No mutations that cause such a defect map to mouse chromosome 11, implying that *l11Jus27* identifies a new gene required for cardiac development¹⁴. *l11Jus15* homozygotes develop a severe oedema, suggesting a circulatory defect (Fig. 2e, f). *l11Jus51* homozygotes exhibit a severe anaemia at birth (Fig. 2g). Just as *Drosophila* lethal mutants have been valuable for dissecting the molecular basis for embryonic patterning¹⁵, these mouse mutations will be valuable for identifying new genes that function in cardiovascular and haematopoietic development.

In addition to the lethal mutants, an additional 24 pedigrees produced phenotypes segregating on mouse chromosome 11 that affect coat, growth/size, skeletal morphology, blood cell abnormalities, or neurological presentation (Table 1; see also Supplementary Table 3). Nine of the mutations isolated on chromosome 11 exhibit severe craniofacial defects, sometimes accompanied by other abnormalities such as seizures and growth defects (craniofacial *crf2*, -6, -5, -8, -12; neurological *nur8*, *l11Jus52*, -53 and -54). With the exception of a craniofacial defect observed in a *Hoxb2* knockout mouse¹⁶, no mouse craniofacial defects have been reported previously on this chromosome. Several human diseases with craniofacial defects map to the region of conserved synteny on human chromosome 17, including Meckel syndrome (OMIM number 24900), lethal microcephaly (OMIM number 607196), mulibrey nanism (OMIM number 253250) and Alexander disease (OMIM number 203450), suggesting that our mutants may be new models of human disease.

A screen for infertility on a subset of 184 pedigrees identified eight pedigrees exhibiting infertile or sub-fertile phenotypes mapping to chromosome 11. Six mutations resulted in male sterility with reduced sperm counts (infertility *inf2*, -4, -5, -6, -8, -9). Histopathology revealed blocks in spermatogenesis in *inf4*, *inf8* and *inf9* males that occurred between round spermatid formation (*inf9*) and late differentiating spermatids (*inf4*; Supplementary Fig. 3). Cysts in multiple organs, including kidney, testis and liver, led to a male sub-fertile phenotype in one mutant, *inf6*, but was not a primary male infertility defect.

The ENU mutation rate in mouse spermatogonial stem cells is one new mutation per locus in every 700 gametes or mutagenized genomes screened¹⁷. Therefore, by screening 735 pedigrees, we have

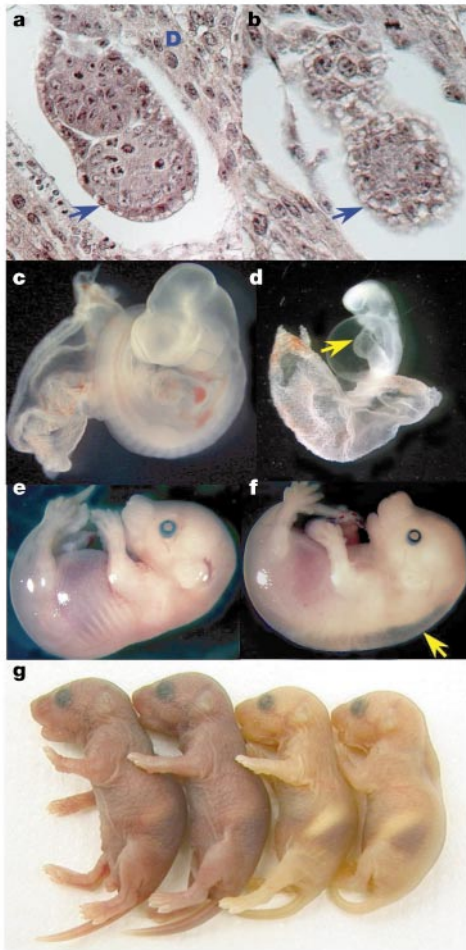


Figure 2 Lethal mutants affect gastrulation and cardiovascular development. **a**, Wild-type embryos at E5.5 develop within the maternal decidua (D), and have well-developed cuboidal epithelium that defines the visceral endoderm (blue arrow). **b**, *l11Jus3* homozygotes have abnormal visceral endoderm and arrest before gastrulation (blue arrow). **c**, Wild-type embryos at E10.0 have well-developed hearts. **d**, *l11Jus27* homozygotes have distended tubular hearts (yellow arrow). **e**, Wild-type embryo at E16.5. **f**, *l11Jus15* homozygotes develop a severe oedema (arrow). **g**, *l11Jus51* homozygotes (right) are pale at birth compared with wild-type littermates (left).

sampled one genome equivalent for this region of mouse chromosome 11. In such a sample, some genes may not have been mutated in any pedigree, whereas others may have been mutated in multiple pedigrees. Consistent with this prediction, obvious visible mutant phenotypes of genes on mouse chromosome 11, such as headless embryos (*Lhx1*)¹⁸, circling behaviour (*Myo15*)¹⁹ and hairless mice (*Foxn1*)²⁰ have not been isolated. In addition, we performed pairwise complementation crosses on mutants in similar phenotypic classes (Supplementary Tables 4 and 5). Of 24 lethal mutations completed, only two (*I11Jus1* and *I11Jus4*) fail to complement each other. These data demonstrate that the great majority of these mutations are in independent genes. It is likely that most of the mutations are in genes not previously mutated or are new alleles of known genes, because most of the developmental and neurological phenotypes do not mimic the phenotypes described in the 45

targeted or spontaneous mutations (Table 1).

Chemical mutagenesis can cause multiple mutations in a single pedigree. The likelihood that linked lethal mutations are being recovered provides an estimate of how commonly linked mutations occur. If we assume a Poisson distribution, a total of 7.4% of chromosome 11 transmissions will carry one lethal mutation, which is not statistically different from the number observed (7.5%; $\chi^2 = 0.008$; $P > 0.90$). The likelihood of a chromosome containing two lethal mutations at the dose of $3 \times 100 \text{ mg kg}^{-1}$ in the C57BL/6J strain is only 3 in 1,000. This is very different from screens in *Drosophila* where over 80% of the chromosomes were predicted to contain one or more lethal mutations when administered a dose of 25 mM ethylmethanesulphonate²¹. The difference can be explained because one whole *Drosophila* chromosome covered by a balancer, such as chromosome 2, contains nearly

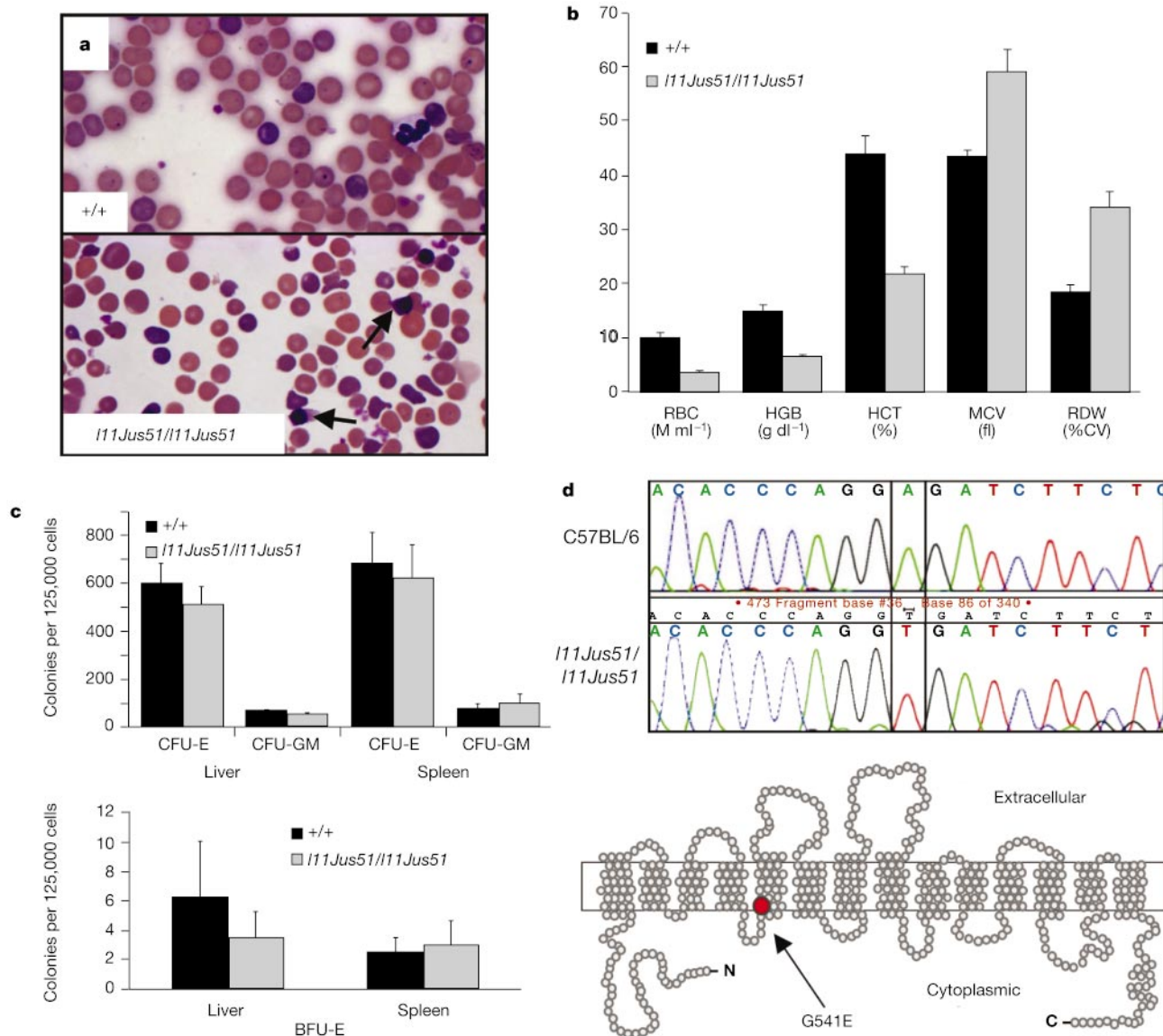


Figure 3 A new *Slc4a1* missense mutation. **a**, Peripheral blood smears from 2-day-old *I11Jus51* homozygotes have nucleated erythroid cells (arrows). **b**, CBC analysis of 8-week-old *I11Jus51* homozygotes shows decreases in erythrocyte numbers (RBC), haemoglobin (HGB) and haematocrit (HCT), with increased mean corpuscular volume (MCV) and red blood cell distribution width (RDW). %CV, % corpuscular volume.

c, Methylcellulose cultures of fetal liver and spleen reveal no differences in progenitor cell numbers and kinetics. CFU-E, colony-forming unit erythroid; CFU-GM, colony forming unit granulocyte/macrophage; BFU-E, burst-forming unit erythroid. **d**, Sequencing *Slc4a1* in *I11Jus51* homozygotes reveals an A-to-T transversion at coding nucleotide 1622, resulting in a glutamic acid to valine change.

35% of the fly genome, whereas our screen targets approximately 2% of the mouse genome. For our screen, the Poisson calculation suggests that the probability of a linked mutation interfering with the phenotype of another is very low.

One benefit of the balancer chromosome is that mutations within the inversion interval are mapped on their isolation, without further crosses, allowing for immediate analysis of candidate genes. For example, *l11Jus51* maps within the balancer interval and 80–90% of homozygous animals die within one week of birth with a severe anaemia (Fig. 2g). Red blood cell morphology is abnormal, and large numbers of nucleated erythroid cells are present in the circulation (Fig. 3a). Rare adult survivors have a macrocytic anaemia, with decreased haemoglobin, haematocrits and red blood cell numbers (Fig. 3b). *In vitro* haematopoietic colony assays of fetal liver and spleen reveal that erythroid progenitor cells are present in normal numbers and possess normal proliferative potential, suggesting that the anaemia is due to abnormal destruction or loss of red blood cells, and not to a defect in production (Fig. 3c). An analysis of the databases shows several genes involved in erythrocyte maturation or maintenance in the balancer interval; most notable is solute carrier protein 4a1 (*Slc4a1*; also known as band 3 or anion exchange protein 1). Two targeted mutations eliminating this protein result in a severe microcytic anaemia with altered red blood cell membrane fragility^{22,23}. We investigated the possibility that *l11Jus51* carried a mutation in *Slc4a1* by performing a cross with the loss of function allele²², which revealed that *l11Jus51* fails to complement the mutation. Subsequent sequencing revealed a missense mutation in the codon for amino acid 541, a residue conserved between mouse and human, changing a glutamic acid to a valine within the fifth transmembrane domain (Fig. 3d). Notably, mutations in *SLC4A1* are associated with hereditary spherocytosis and resistance to malarial infection in humans²⁴. Many of the missense disease mutations in human *SLC4A1* cause an autosomal dominant form of the disease that results in renal tubular acidosis^{25,26}. *l11Jus51* is the first mouse missense mutation in *Slc4a1*, and causes a macrocytic anaemia rather than a microcytic anaemia as in the deletion alleles, suggesting that it will be useful for dissecting the biochemical basis of the loss of protein function without its complete absence. In addition, preliminary histopathology reveals kidney abnormalities; therefore, *l11Jus51* may prove valuable as a mouse model for renal tubular acidosis.

New mouse mutations are the key to functional genomic studies that will unravel mammalian systems biology. Genome sequencing shows that the mammalian genome contains fewer genes than previously predicted¹. However, transcriptome and proteome analysis suggests that higher organism complexity is produced not by an increase in the absolute number of genes but by changes in temporal and spatial transcript expression, as well as alternative transcript and protein isoform production²⁷. Because ENU mutagenesis is phenotype-driven, it is a powerful way to discover gene function because no prior assumptions are made regarding a particular transcript or protein product. Although the *Trp53–Wnt3* interval of mouse chromosome 11 is predicted to contain over 700 genes and already has a large number of knockout mutations²⁸, we have tripled the number of mutants in this region with a one-genome coverage screen. The recovery of 142 mutations on other mouse chromosomes demonstrates that the balancer mutagenesis strategy has multiple benefits for the efficient functional annotation of the mammalian genome. □

Methods

Mouse strains

Inv(11)8Brd^{1Trp53–Wnt3} was engineered in 129/SvEv embryonic stem cells, and is maintained by subsequent backcrosses onto the inbred strains C57BL/6J (11 generations), 129S6/SvEvTac (8 generations) and C3HeB/FeJ (6 generations). Most pedigrees screened for mutations used 129.Inv(11)8Brd. The *Rex* mutation was acquired from The Jackson

Laboratory as the RSV/Le stock, and was subsequently backcrossed to make it congenic ($N > 10$) on the same three strains.

Mutagenesis and screening

C57BL/6J males (Jackson Laboratory) were injected intraperitoneally at the age of 8–10 weeks with 100 mg ENU per kg body weight weekly for three weeks³. Forty-seven fertile treated males were mated as in Fig. 1b. Of 905 F₁ animals, 735 produced enough litters at the third (F₃) generation to test that pedigree for phenotypes. A pedigree was bred as a potential new mutation if: (1) no test class animals were observed in a total of 24 carrier animals ($\chi^2 = 10.5$; $P < 0.01$); (2) all test class animals exhibited a similar phenotype; or (3) at least two or more animals of test or carrier class exhibited the same phenotype in a single pedigree. Additional phenotypes were found by letting the animals age to 3 months, performing complete blood counts (CBCs) with a differential analysis of white blood cells on 517 pedigrees²⁹, or testing for infertility in 184 pedigrees. Additional pedigrees were screened for biochemical, bone or soft tissue abnormalities (see Supplementary Information for phenotyping details).

Characterization of lethal mutants

If a pedigree produced no test class animals in a cohort of at least 24 carriers, the carrier class animals were used to generate a stock of mice. To eliminate any interfering phenotypes or accessory mutations in the pedigree, carrier animals were out crossed at least once to C3H.Inv(11)8Brd/Re animals, and then Inv/m animals were intercrossed to recover the original mutation. To determine the stage of death, timed matings were performed on carrier females by examining them each morning for a vaginal plug. Evidence of a plug was designated embryonic day (E)0.5, and females were killed for embryo examination at various stages. Mutant embryos were fixed in Bouin's fixative, embedded in paraffin, sectioned at 5 μ m, and stained with haematoxylin and eosin as previously described³⁰. Using PCR, embryos were genotyped for the presence of the human HPRT cassette present in the engineered inversion. DNA was prepared from embryonic tail or yolk sac samples and PCR was performed as a multiplex reaction with primers to the HPRT mini gene to detect the inversion chromosome and primers to mouse BAC clone RPC123–132B20 T7 end as a control for the presence of DNA. Conditions and primer sequences are available on request.

Sequencing

After PCR amplification, genomic DNA prepared from tails was sequenced directly using BigDye Terminators v3.0 (Applied Biosystems) according to the manufacturer's instructions. All *Slc4a1* exons and splice junctions were sequenced in their entirety in multiple homozygous *l11Jus51* animals and the treated strain C57BL/6J; oligonucleotide primer sequences are available on request. No differences between *l11Jus51* and C57BL/6J sequences were detected other than the A-to-T transversion at coding nucleotide 1622, which abolishes a *Bgl*III endonuclease restriction site.

Statistics and web display

Poisson distribution calculations were based on the observation of 55 lethal mutations in 735 pedigrees. Thus, $P(0) = e^{-m} = 0.925$, $m = 0.08$; $P(1) = 0.074$; $P(2) = 0.003$; $P(3) = 0.00008$. χ^2 tests are calculated with Yates' correction. See <http://www.mouse-genome.bcm.tmc.edu> (linked to a relational database) for phenotype information on the mutants generated from this mutagenesis screen. The interface to retrieve data from the SQL server was developed using Active Server Pages (ASP). Users can query mutant and map location, retrieve phenotype information and request mutants from this web page. An image management system, ImageView, allows public viewing of images. Movies can be viewed after registration on the web page.

Received 26 March; accepted 10 June 2003; doi:10.1038/nature01865.

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Justice, M. J. in *Genetics and Transgenics: A Practical Approach* (eds Jackson, I. & Abbott, C.) 185–215 (Oxford Univ. Press, Oxford, 1999).
4. Zheng, B. *et al.* Engineering a balancer chromosome in the mouse. *Nature Genet.* **22**, 375–378 (1999).
5. Zheng, B., Mills, A. A. & Bradley, A. A system for rapid generation of coat color-tagged knockouts and defined chromosomal rearrangements in mice. *Nucleic Acids Res.* **27**, 2354–2360 (1999).
6. Muller, H. J. Genetic variability, twin hybrids, and constant hybrids, in a case of balanced lethal factors. *Genetics* **3**, 422–499 (1918).
7. Kasarskis, A., Manova, K. & Anderson, K. V. A phenotype-based screen for embryonic lethal mutations in the mouse. *Proc. Natl Acad. Sci. USA* **95**, 7485–7490 (1998).
8. Hentges, K., Thompson, K. & Peterson, A. The flat-top gene is required for the expansion and regionalization of the telencephalic primordium. *Development* **126**, 1601–1609 (1999).
9. Herron, B. J. *et al.* Efficient generation and mapping of recessive developmental mutations using ENU mutagenesis. *Nature Genet.* **30**, 185–189 (2002).
10. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
11. Miklos, G. L. & Rubin, G. M. The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**, 521–529 (1996).
12. Shedlovsky, A., King, T. R. & Dove, W. F. Saturation germ line mutagenesis of the murine *t* region including a lethal allele at the *quaking* locus. *Proc. Natl Acad. Sci. USA* **85**, 180–184 (1988).
13. Rinchik, E. M. & Carpenter, D. A. N-ethyl-N-nitrosourea mutagenesis of a 6- to 11-cM subregion of the *Fah-Hbb* interval of mouse chromosome 7: completed testing of 4,557 gametes and deletion mapping and complementation analysis of 31 mutations. *Genetics* **152**, 373–383 (1999).

14. Conway, S. J., Kruzynska-Frejtag, A., Kneer, P. L., Machnicki, M. & Koushik, S. V. What cardiovascular defect does my prenatal mouse mutant have, and why? *Genesis* **35**, 1–21 (2003).
15. Nusslein-Volhard, C., Frohnhof, H. G. & Lehmann, R. Determination of anteroposterior polarity in *Drosophila*. *Science* **238**, 1675–1681 (1987).
16. Barrow, J. R. & Capecchi, M. R. Targeted disruption of the *Hoxb-2* locus in mice interferes with expression of *Hoxb-1* and *Hoxb-4*. *Development* **122**, 3817–3828 (1996).
17. Hitotsumachi, S., Carpenter, D. A. & Russell, W. L. Dose-repetition increases the mutagenic effectiveness of *N*-ethyl-*N*-nitrosourea in mouse spermatogonia. *Proc. Natl Acad. Sci. USA* **82**, 6619–6621 (1985).
18. Shawlot, W. & Behringer, R. R. Requirement for *Lim1* in head-organizer function. *Nature* **374**, 425–430 (1995).
19. Probst, F. J. et al. Correction of deafness in *shaker-2* mice by an unconventional myosin in a BAC transgene. *Science* **280**, 1444–1447 (1998).
20. Nehls, M., Pfeifer, D., Schorpp, M., Hedrich, H. & Boehm, T. New member of the winged-helix protein family disrupted in mouse and rat *nude* mutations. *Nature* **372**, 103–107 (1994).
21. Nusslein-Volhard, C., Wieschaus, E. & Kluding, H. Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster* I. Zygotic loci on the second chromosome. *Roux Arch. Dev. Biol.* **193**, 267–282 (1984).
22. Peters, L. L. et al. Anion exchanger 1 (band 3) is required to prevent erythrocyte membrane surface loss but not to form the membrane skeleton. *Cell* **86**, 917–927 (1996).
23. Southgate, C. D., Chishti, A. H., Mitchell, B., Yi, S. J. & Palek, J. Targeted disruption of the murine erythroid band 3 gene results in spherocytosis and severe haemolytic anaemia despite a normal membrane skeleton. *Nature Genet.* **14**, 227–230 (1996).
24. Jarolim, P. et al. Mutations of conserved arginines in the membrane domain of erythroid band 3 lead to a decrease in membrane-associated band 3 and to the phenotype of hereditary spherocytosis. *Blood* **85**, 634–640 (1995).
25. Karet, F. E. et al. Mutations in the chloride-bicarbonate exchanger gene AE1 cause autosomal dominant but not autosomal recessive distal renal tubular acidosis. *Proc. Natl Acad. Sci. USA* **95**, 6337–6342 (1998).
26. Bruce, L. J. et al. Familial distal renal tubular acidosis is associated with mutations in the red cell anion exchanger (band 3, AE1) gene. *J. Clin. Invest.* **100**, 1693–1707 (1997).
27. The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
28. Behringer, R. R. (ed.) Mouse knockout and mutation database. *BioMedNet* (<http://research.bmn.com/mkmd>) (2003).
29. Kile, B. T., Mason-Garrison, C. & Justice, M. J. Sex and strain-related differences in the peripheral blood cell values of inbred mouse strains. *Mamm. Genome* **14**, 81–85 (2003).
30. Noveroske, J. N. et al. *Quaking* is essential for blood vessel development. *Genesis* **32**, 218–230 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Viator, C. Dinh, S. Moncrief, A. Zalud, J. Maffucci, C. Mason-Garrison, K. Firozi, M. Alviento, C. Hubbard, B. Hasson and M. Scantlin for technical assistance, and J. Zhong and M. Patterson for database support. We also thank L. Peters for the gift of *Slc4a1* knockout mice. Y. Furuta, H. Bellen, S. Lovell, S. Watowich and H. Gilbert are thanked for critical reading of this manuscript. This work was supported by NIH grants to M.J.J. and A.B. K.E.H. was supported by an NIH-NRSA grant. B.T.K. is a Fellow of the Leukemia Research Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.J. (mjustice@bcm.tmc.edu).

The plastid *clpP1* protease gene is essential for plant development

Hiroshi Kuroda¹ & Pal Maliga^{1,2}

¹Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, 190 Frelinghuysen Road, Piscataway, New Jersey 08854-8020, USA

²Department of Plant Biology, Rutgers, The State University of New Jersey, 59 Dudley Road, New Brunswick, New Jersey 08901, USA

Plastids of higher plants are semi-autonomous cellular organelles that have their own genome and transcription-translation machinery¹. Examples of plastid functions are photosynthesis and biosynthesis of starch, amino acids, lipids and pigments². Plastid functions are encoded in ~120 plastid genes³ and ~3,000 nuclear genes^{2,3}. Although many embryo and seedling lethal

nuclear genes are required for chloroplast biogenesis^{4–6}, until now deletion of plastid genes either had no phenotypic consequence (8 genes), or caused a mutant phenotype but did not affect viability (13 genes)^{7–10}. Here we identify an essential plastid gene. By using the CRE-lox site-specific recombination system^{11,12} we have deleted *clpP1* (caseinolytic protease P1), one of the three genes (*clpP1*, *ycf1* and *ycf2*) whose disruption had previously only been possible in a fraction of the 1,000–10,000 plastid genome copies in a cell^{7,13}. Loss of the *clpP1* gene product, the ClpP1 protease subunit¹⁴, results in ablation of the shoot system of tobacco plants, suggesting that ClpP1-mediated protein degradation is essential for shoot development.

CRE, a site-specific recombinase derived from the P1 bacteriophage, excises any DNA sequence between two directly oriented 34-base-pair *lox* sites¹⁵. The *clpP1* exon 2 (E2) and exon 3 (E3) were first bracketed by two directly oriented *lox* sites in the plastid genome (Fig. 1a). The *clpP1* E2 and E3 encode conserved amino-acid residues essential for ClpP1 catalytic function^{14,16}. A *Cre* gene encoding a plastid-targeted CRE enzyme was then introduced into the nucleus by a sexual cross with a plant carrying a nuclear *Cre* gene. CRE, translated in the cytoplasm, was imported into all plastids and excised the *clpP1* segment between the two *lox* sites (Fig. 1b).

Tobacco plants with engineered *clpP1* genes—now flanked by *lox* sites ('floxed'), and denoted *clpP1*^{fl}—were obtained by plastid transformation. Plastid transformation was carried out with vector pHK85, in which the engineered *clpP1*^{fl} was linked with a spectinomycin resistance (*aadA*) marker gene (Fig. 2a). The transforming DNA was introduced into chloroplasts by bombardment of tobacco leaves, and transplastomic clones were selected by spectinomycin resistance. Transformed plastid genomes were obtained by replacing the plastid *clpP1* gene with the engineered *clpP1*^{fl} and *aadA* genes carried in the vector (Figs 1a and 2a). Uniform transformation of plastid genomes was confirmed by DNA gel blot analysis (lane *clpP1*^{fl}, Fig. 3a)¹⁷. Plants regenerated from the transformed tissue were indistinguishable from wild-type plants. Thus, introduction of *lox* sites and *aadA* in the *clpP1* operon did not interfere with plastid function.

Tobacco plants carrying a nuclear *Cre* gene were obtained by *Agrobacterium*-mediated transformation. The *Agrobacterium* vectors carried a *Cre* gene engineered for expression in the plant nucleus; they had a constitutive P2' promoter and a *nos* sequence encoding a polyadenylation site. The encoded CRE protein is plastid targeted because the *Cre* coding region is translationally fused at its amino terminus with a DNA segment encoding the Rubisco small subunit transit peptide (Fig. 2b).

To trigger excision of *clpP1*^{fl}, plants with engineered plastids were pollinated with nuclear *Cre* lines. Note that in tobacco only the maternal parent transmits plastids to the seed progeny, thus the seedlings carried only engineered plastids containing *clpP1*^{fl}. The *Cre* lines were maintained as heterozygotes, therefore pollination of *clpP1*^{fl} plants (maternal parent) with nuclear *Cre* lines was expected to yield seed progeny in which about half of the seedlings carried the nuclear *Cre*. It is well known that the developmental timing and level of transgene expression in the nucleus is modulated by host genome sequences adjacent to the insertion site¹⁸. Thus, we expected differences in CRE activity in the seed progeny derived from crosses with independently transformed *Cre* lines.

To evaluate the consequence of *clpP1*^{fl} excision, seed progeny derived from the crosses were germinated in sterile culture in the absence of antibiotics. Under this condition, wild-type seedlings are green (Fig. 4f). The cross of *clpP1*^{fl} plants as maternal parent and Nt-*Cre30B* as pollen parent¹⁹ segregated green and white progeny (302 to 317; ~1:1) (Fig. 4a). The green seedlings developed into normal plants, whereas the white seedlings failed to develop a shoot system even after ~6 months (182 days; Fig. 4b). Polymerase chain reaction (PCR) analysis confirmed that the seedlings with white

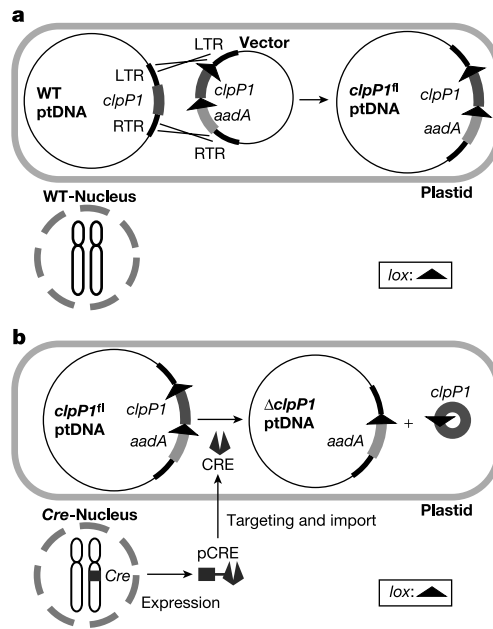


Figure 1 Strategy for deletion of the plastid *clpP1* gene by the CRE recombinase. **a**, The *clpP1* gene in the plastid transformation vector (flanked by *lox* sites, *clpP1*^{fl}) replaces *clpP1* in the plastid genome (wild-type (WT) ptDNA) by two homologous recombination events. In the vector, *clpP1* is linked to *aadA*. Left and right targeting regions (LTR and

RTR, respectively) are plastid DNA sequences. **b**, CRE encoded in a nuclear *Cre* gene is imported into plastids and excises *clpP1*^{fl} to yield $\Delta clpP1$ ptDNA. pCRE is the CRE precursor protein, from which the plastid-targeting N-terminal extension is cleaved off upon import to yield CRE.

cotyledons carried the *Cre* gene (Fig. 3a, *Cre*). Excision of *clpP1* was tracked in total cellular DNA extracted from cotyledons, hypocotyls and roots of individual, 9-day-old seedlings. DNA gel blot analysis did not detect *clpP1* copies in white cotyledons and hypocotyls (Fig. 3a), indicating efficient excision of the *clpP1* segment from the

plastid genome (although *clpP1* copies could still be detected by the more sensitive PCR assay). Reduction of ClpP1 protein levels to ~50% already caused a pale green seedling phenotype²⁰. In the white *clpP1*^{fl} $\times$ Nt-Cre30B seedlings, intact *clpP1* copy number was reduced to <1%, with <5% of wild-type messenger RNA present

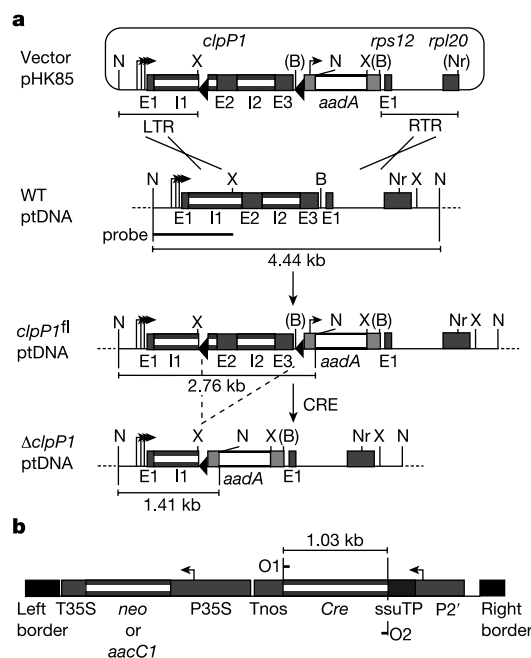


Figure 2 Maps of transformation vectors and plastid DNA. **a**, Plasmid transformation vector pHK85, wild-type plastid DNA (WT ptDNA), transplastomic *clpP1*^{fl} ptDNA and *clpP1* deletion derivative ($\Delta clpP1$ ptDNA). Shown are: transcription initiation sites, horizontal arrows; plastid genes *clpP1*, *rps12*, *rpl20*; selectable marker, *aadA*; *lox* sites, triangles; exon, E; intron, I; left and right targeting regions, LTR and RTR; restriction sites *Bst*BI (B), *Nco*I (N), *Nru*I (Nr) and *Xba*I (X); eliminated restriction sites are in parentheses. **b**, T-DNA

region of *Agrobacterium* binary vectors. Shown are the *Cre* gene, the kanamycin resistance (*neo*; pK027)¹¹ or gentamycin resistance (*aacC1*; pK030)¹⁹ marker gene, and left and right border sequences. The *Cre* gene has the P2' promoter; the Rubisco small subunit transit peptide (ssuTP) and the *nos* terminator (Tnos). The marker genes are expressed in the P35S-T35S cassette. O1 and O2 are PCR primers.

(not shown) and no detectable ClpP1 protein (<2%; Fig. 5), yielding a *clpP1* knockout phenotype. A small fraction (<10%) of plastid genomes in the roots, however, still carried an intact *clpP1*, probably owing to relatively low CRE expression in roots (Fig. 3a, Root).

Seed progeny derived from the cross between *clpP1*^{fl} plants as maternal parent and three other nuclear *Cre* lines (Nt-*Cre*1-100, Nt-*Cre*2-100, Nt-*Cre*2-200)¹¹ as pollen parents developed green shoots (Fig. 4c–e), even though about 50% of the seedlings carried the *Cre* (Fig. 3b–d). CRE protein levels in the cotyledons of these lines were at least fourfold lower than in Nt-*Cre*30B cotyledons (not shown). Absence of white seedlings in the progeny of the three additional *Cre* lines was due to inefficient excision of *clpP1*, with *clpP1* deletion in 5% to 50% of plastid genome copies in seedling cotyledons of the Nt-*Cre*1-100 (Fig. 3b) and Nt-*Cre*2-200 (Fig. 3d) progeny. Protein output from the reduced number of *clpP1* copies in the green seedlings was apparently compensated by expression of intact *clpP1* genes (Fig. 5) and these seedlings developed a shoot system. In the seed progeny of the Nt-*Cre*2-100 line, with >50% of plastid genome copies lacking *clpP1* (Fig. 3c), a pale green mottling was observed, suggesting localized reduction in ClpP1 levels (Fig. 4d).

Progeny obtained by self-pollination of nuclear *Cre* lines and progeny derived from crosses using a wild-type plant as a maternal parent were green. Seed progeny derived from a control cross in which Nt-*Cre*30B was the pollen parent and a transplastomic plant with only one *lox* site downstream of *clpP1* was the maternal parent (and therefore lacking directly oriented *lox* sites for *clpP1* excision),

was also green (data not shown). Thus, white cotyledons and an arrest of shoot development were obtained only in crosses with the Nt-*Cre*30B line, in which *Cre* expression in the nucleus could be linked to efficient *clpP1* excision and a reduction in ClpP1 protein levels.

The polyploid nature of the plastid genome, which may be present in 1,000 to 10,000 copies in a cell, complicates identification of essential plastid genes. We have shown here that CRE-mediated excision is efficient enough to study plastid gene function by excision of plastid genes in somatic cells. Data presented here settle a long-standing argument as to whether or not plastid genes are essential for plant development and viability. Plastid ribosome-deficient plants have been described in oilseed rape (*Brassica napus*)²¹, maize^{22,23} and barley²⁴. These plants lack the capacity to translate any plastid mRNA, thus apparently no plastid-encoded proteins are essential in the absence of plastid protein synthesis. The *clpP1* gene is non-essential in non-photosynthetic cultured cells, as suggested by its absence in some maize cell culture lines²⁵. However, *clpP1* is retained in a non-photosynthetic, metabolically active parasitic plant in which most photosynthetic plastid genes have been eliminated²⁶. Furthermore, attempts to obtain homoplasmic *clpP1* tobacco knockout lines have failed¹³. Data presented here indicate that the *clpP1* gene product is essential for the execution of the normal shoot developmental programme in tobacco seedlings.

The Clp protease is a two-component enzyme. In *Escherichia coli* it consists of an endopeptidase, ClpP, that relies on the unfolding activity of a molecular chaperone, either ClpA or ClpX. ClpA or ClpX flanks both ends of the proteolytic component, providing

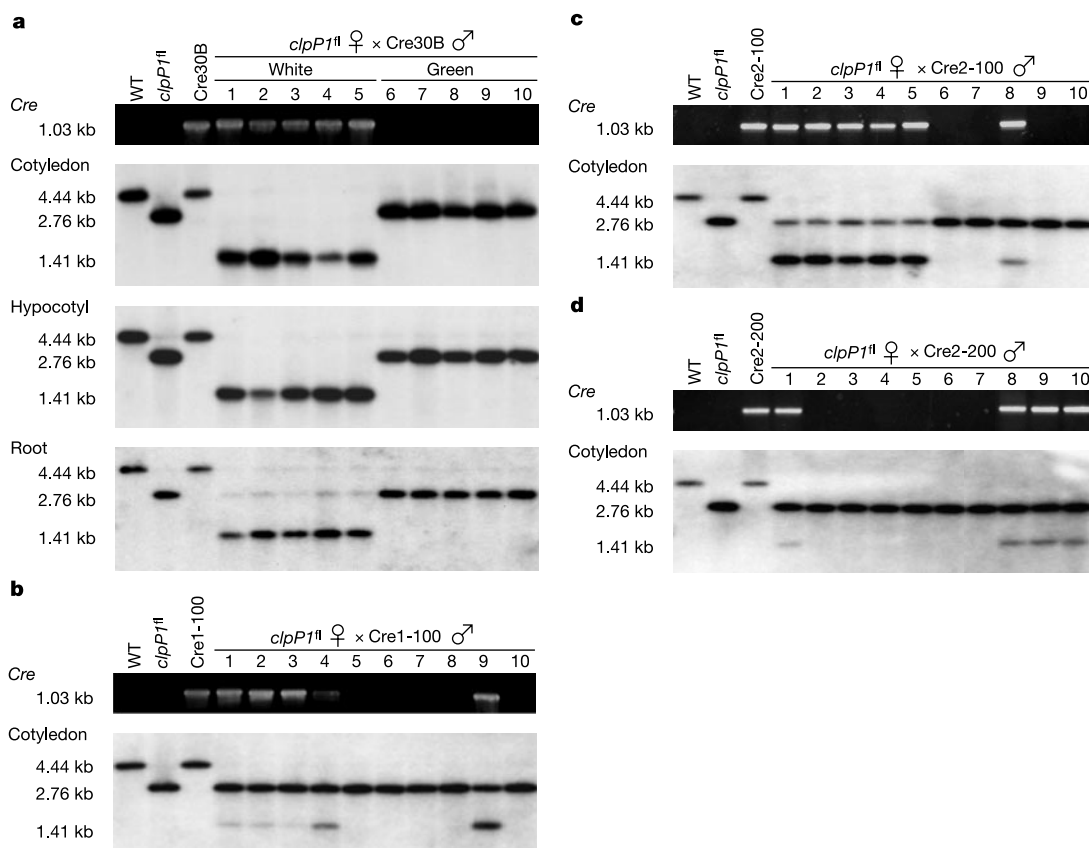


Figure 3 CRE-mediated excision of *clpP1* in the seed progeny. DNA was tested from individual seedlings derived from the cross of **a**, Nt-pHK85 *clpP1*^{fl} (maternal parent) and as pollen parents the Nt-*Cre*30B; **b**, Nt-*Cre*1-100; **c**, Nt-*Cre*2-100; or **d**, the Nt-*Cre*2-200 lines. Seedlings were tested for *Cre* by PCR analysis¹¹ (top). Excision of *clpP1* was tested

by DNA gel blot analysis using a ³²P-labelled *Nco*I-*Xba*I *clpP1* probe (nucleotides 73,739–74,956; GenBank Z00044; Fig. 2a). 300 ng *Nco*I-digested DNA was loaded per lane.

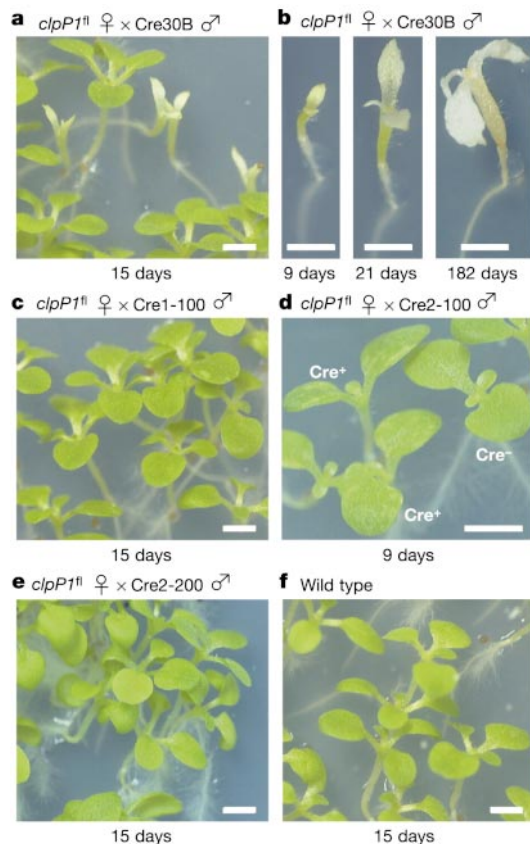


Figure 4 Seedling phenotypes in crosses between Nt-pHK85 *clpP1*^{fl} plants as maternal parent and nuclear *Cre* pollen parents. The seedlings were germinated in the absence of antibiotics on sucrose-containing medium. Shown are progeny of crosses: **a, b**, *clpP1*^{fl} × Nt-*Cre30B*; **c**, *clpP1*^{fl} × Nt-*Cre1-100*; **d**, *clpP1*^{fl} × Nt-*Cre2-100*; **e**, *clpP1*^{fl} × Nt-*Cre2-200*; **f**, wild type. Scale bars, 3 mm.

gateways to the proteolytic sites. The regulatory chaperones confer substrate specificity and deliver the unfolded protein into the ClpP proteolytic chamber^{27–29}. Within the chloroplast, the Clp protease is thought to be responsible for the majority of protein degradation. In *Arabidopsis* plastids, in addition to the plastid-encoded proteolytic ClpP1, there are at least four proteolytic and two regulatory Clp subunits encoded by nuclear genes¹⁴. Lack of shoot development in the *clpP1* deletion seedlings suggests that the nuclear-encoded catalytic subunit genes in tobacco cannot replace the plastid *clpP1* gene. Thus, inhibition of shoot development may be due to insufficient expression of catalytic subunits from nuclear genes in the developing seedlings. An alternative, more intriguing, possibility is the lack of degradation of a regulatory protein that is a specific substrate for the ClpP1 isoform. □

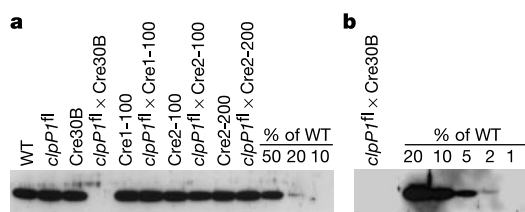


Figure 5 Immunoblot to detect accumulation of ClpP1 in seedling cotyledons²⁰. **a**, 35 µg or **b**, 200 µg total soluble protein was loaded per lane. Wild-type protein extract serves as a reference for quantification.

Methods

Plastid vector pHK85 (pUC120 plasmid derivative; Fig. 2a) carries sequences between nucleotides 71,201 and 74,956 of the tobacco plastid genome (GenBank Z00044). One *lox* site was inserted between nucleotides 73,738 and 73,739 in *clpP1* intron 1. A second *lox* site and the spectinomycin-resistance (*aadA*) gene were inserted at the *Bst*BI site (nucleotide 72,423) between *clpP1* exon 3 and *rps12* exon 1. The *aadA* gene was expressed in a *psbA* cassette derived from plasmid pJS25 (ref. 30).

Received 29 April; accepted 11 July 2003; doi:10.1038/nature01909.

1. Sugiura, M. The chloroplast genome. *Plant Mol. Biol.* **19**, 149–168 (1992).
2. Leister, D. Chloroplast research in the genomic era. *Trends Genet.* **19**, 47–56 (2003).
3. Martin, W. *et al.* Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA* **99**, 12246–12251 (2002).
4. McElver, J. *et al.* Insertional mutagenesis of genes required for seed development in *Arabidopsis thaliana*. *Genetics* **159**, 1751–1763 (2001).
5. Budziszewski, G. J. *et al.* *Arabidopsis* genes essential for seedling viability: Isolation of insertional mutants and molecular cloning. *Genetics* **159**, 1765–1778 (2001).
6. Maréchal, E. Chloroplast biogenesis and function are first in the list of essential *Arabidopsis* genes. *Trends Plant Sci.* **7**, 99–100 (2002).
7. Bock, R. in *Progress in Botany* (eds Esser, K., Lüttge, U., Beyschlag, W. & Hellwig, F.) 106–131 (Springer, Berlin, 2002).
8. Hager, M., Hermann, M., Biehler, K., Krieger-Liszka, A. & Bock, R. Lack of the small plastid-encoded PsbJ polypeptide results in a defective water-splitting apparatus of photosystem II, reduced photosystem I levels, and hypersensitivity to light. *J. Biol. Chem.* **277**, 14031–14039 (2002).
9. Swiatek, M. *et al.* Effects of selective inactivation of individual genes for low-molecular-mass subunits on the assembly of photosystem II, as revealed by chloroplast transformation: The *psbEFLJ* operon in *Nicotiana tabacum*. *Mol. Genet. Genomics* **268**, 699–710 (2003).
10. Swiatek, M. *et al.* PCR analysis of pulse-field gel electrophoresis-purified plastid DNA, a sensitive tool to judge the hetero-/homoplasmic status of plastid transformants. *Curr. Genet.* **43**, 45–53 (2003).
11. Corneille, S., Lutz, K., Svab, Z. & Maliga, P. Efficient elimination of selectable marker genes from the plastid genome by the CRE-*lox* site-specific recombination system. *Plant J.* **72**, 171–178 (2001).
12. Hajdukiewicz, P. T. J., Gilbertson, L. & Staub, J. M. Multiple pathways for *Cre/lox*-mediated recombination in plastids. *Plant J.* **27**, 161–170 (2001).
13. Shikanai, T. *et al.* The chloroplast *clpP* gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco. *Plant Cell Physiol.* **42**, 264–273 (2001).
14. Adam, Z. & Clarke, A. K. Cutting edge of chloroplast proteolysis. *Trends Plant Sci.* **7**, 451–456 (2002).
15. Ow, D. W. Recombinase-directed plant transformation for the post-genomic era. *Plant Mol. Biol.* **48**, 183–200 (2002).
16. Peltier, J. B., Ytterberg, J., Liberles, D. A., Roepstroff, P. & van Wijk, K. J. Identification of a 350 kDa ClpP protease complex with 10 different Clp isoforms in chloroplasts of *Arabidopsis thaliana*. *J. Biol. Chem.* **276**, 16318–16327 (2001).
17. Svab, Z. & Maliga, P. High-frequency plastid transformation in tobacco by selection for a chimeric *aadA* gene. *Proc. Natl Acad. Sci. USA* **90**, 913–917 (1993).
18. Springer, P. S. Gene traps: Tools for plant development and genomics. *Plant Cell* **12**, 1007–1020 (2000).
19. Corneille, S., Lutz, K. A., Azhagiri, A. K. & Maliga, P. Identification of functional *lox* sites in the plastid genome. *Plant J.* (in the press).
20. Kuroda, H. & Maliga, P. Overexpression of the *clpP* 5′-untranslated region in a chimeric context causes a mutant phenotype, suggesting competition for a *clpP*-specific RNA maturation factor in tobacco chloroplasts. *Plant Physiol.* **129**, 1600–1606 (2002).
21. Zubko, M. K. & Day, A. Stable albinism induced without mutagenesis: A model for ribosome-free plastid inheritance. *Plant J.* **15**, 265–271 (1998).
22. Walbot, V. & Coe, E. H. J. Nuclear gene *iojap* conditions a programmed change to ribosome-less plastids in *Zea mays*. *Proc. Natl Acad. Sci. USA* **76**, 2760–2764 (1979).
23. Han, C. D., Coe, E. H. J. & Martienssen, R. A. Molecular cloning and characterization of the maize *iojap* (*ij*), a pattern striping in maize. *EMBO J.* **11**, 4037–4046 (1992).
24. Hess, W. R., Prombona, A., Fieder, B., Subramanian, A. R. & Börner, T. Chloroplast *rps15* and the *rpoB/C1/C2* gene cluster are strongly transcribed in ribosome-deficient plastids: Evidence for a functioning non-chloroplast-encoded RNA polymerase. *EMBO J.* **12**, 563–571 (1993).
25. Cahoon, A. B., Cunningham, K. A. & Stern, D. B. The plastid *clpP* gene may not be essential for plant cell viability. *Plant Cell Physiol.* **44**, 93–95 (2003).
26. Wolfe, K. H., Morden, C. W. & Palmer, J. D. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl Acad. Sci. USA* **89**, 10648–10652 (1992).
27. Wickner, S., Maurizi, M. R. & Gottesman, S. Posttranslational quality control: Folding, refolding, and degrading proteins. *Science* **286**, 1888–1893 (1999).
28. Flynn, J. M., Neher, S. B., Kim, Y.-I., Sauer, R. T. & Baker, T. A. Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. *Mol. Cell* **11**, 671–683 (2003).
29. Zhou, Y., Gottesman, S., Hoskins, J. R., Maurizi, M. R. & Wickner, S. The RssB response regulator directly targets σ^S for degradation by ClpXP. *Genes Dev.* **15**, 627–637 (2001).
30. Staub, J. M. & Maliga, P. Accumulation of D1 polypeptide in tobacco plastids is regulated via the untranslated region of the *psbA* mRNA. *EMBO J.* **12**, 601–606 (1993).

Acknowledgements We thank K. Lutz and S. Corneille for the nuclear *Cre* plants, and Z. Adam for the ClpP1 antibody. This research was supported by a Rutgers F&A special project grant.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.M. (maliga@waksman.rutgers.edu).

Oleyethanolamide regulates feeding and body weight through activation of the nuclear receptor PPAR- α

Jin Fu¹, Silvana Gaetani¹, Fariba Oveisi¹, Jesse Lo Verme¹, Antonia Serrano², Fernando Rodríguez de Fonseca², Anja Rosengarth³, Hartmut Luecke³, Barbara Di Giacomo⁴, Giorgio Tarzia⁴ & Daniele Piomelli¹

¹Department of Pharmacology, University of California, Irvine, California 92697-4625, USA

²Fundación Hospital Carlos Haya, Malaga, Spain

³Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697-4625, USA

⁴Institute of Medicinal Chemistry, University of Urbino, Italy

Oleyethanolamide (OEA) is a naturally occurring lipid that regulates satiety and body weight^{1,2}. Although structurally related to the endogenous cannabinoid anandamide, OEA does not bind to cannabinoid receptors and its molecular targets have not been defined. Here we show that OEA binds with high affinity to the peroxisome-proliferator-activated receptor- α (PPAR- α), a nuclear receptor that regulates several aspects of lipid metabolism. Administration of OEA produces satiety and reduces body weight gain in wild-type mice, but not in mice deficient in PPAR- α . Two distinct PPAR- α agonists have similar effects that are also contingent on PPAR- α expression, whereas potent and selective agonists for PPAR- γ and PPAR- β/δ are ineffective. In the small intestine of wild-type but not PPAR- α -null mice, OEA regulates the expression of several PPAR- α target genes: it initiates the transcription of proteins involved in lipid metabolism and represses inducible nitric oxide synthase, an enzyme that may contribute to feeding stimulation. Our results, which show that OEA induces satiety by activating PPAR- α , identify an

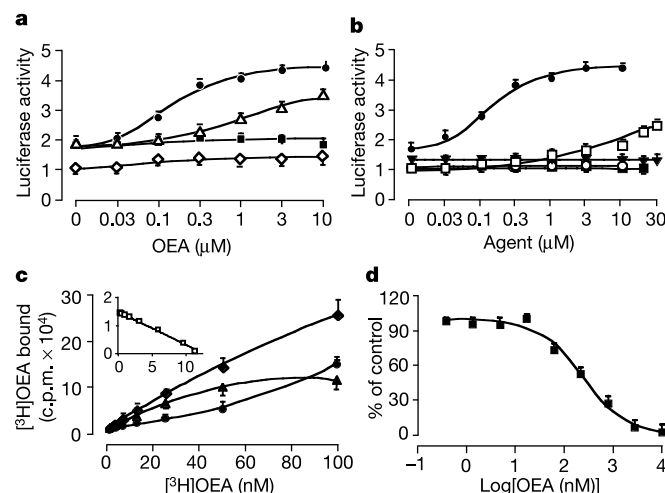


Figure 1 OEA is a high-affinity PPAR- α agonist. **a**, **b**, Activation of human PPAR- α by OEA. **a**, OEA activates human PPAR- α (filled circles) and PPAR- β/δ (open triangles), but not PPAR- γ (filled squares) or RXR (open diamonds). **b**, Effects of OEA (filled circles), oleic acid (open squares), stearyl ethanolamide (filled triangles), myristyl ethanolamide (filled squares) and anandamide (open circles) on PPAR- α activation. **c**, Saturation binding of [³H]OEA to purified mouse PPAR- α LBD. Shown are total binding (filled diamonds), nonspecific binding (filled circles) and specific binding (filled triangles). The inset shows a Scatchard transformation (bound/[bound + free]) of the specific binding data. **d**, Competition binding of OEA to purified mouse PPAR- α LBD. Results are the mean $\pm$ s.e.m. of 16 (**a**, **b**) or 10 (**c**, **d**) experiments.

unexpected role for this nuclear receptor in regulating behaviour, and raise possibilities for the treatment of eating disorders.

Because OEA resembles natural compounds that activate PPARs, such as oleic acid^{3–5}, we considered that it may interact with a member of this nuclear receptor family^{6–8}. To test this idea, we engineered HeLa cells to express a luciferase reporter gene together with the ligand-binding domain (LBD) of human PPAR- α , PPAR- β/δ , PPAR- γ or retinoid-X receptor (RXR) fused to the yeast GAL4 DNA-binding domain⁹. In transactivation assays, OEA activated PPAR- α with a half-maximal concentration (EC_{50}) of 120 ± 1 nM and activated PPAR- β/δ with an EC_{50} of 1.1 ± 0.1 μ M ($n = 16$), whereas it had no effect on PPAR- γ or RXR (Fig. 1a).

To study the structural selectivity of PPAR- α activation, we tested several OEA analogues. As previously reported^{3–5}, oleic acid activated the receptor with micromolar potency ($EC_{50} = 10.3 \pm 0.21$ μ M; $n = 16$), whereas stearyl ethanolamide, myristyl ethanolamide and anandamide¹⁰ had no effect (Fig. 1b). The synthetic agonists Wy-14643 (ref. 11) and GW7647 (ref. 12)

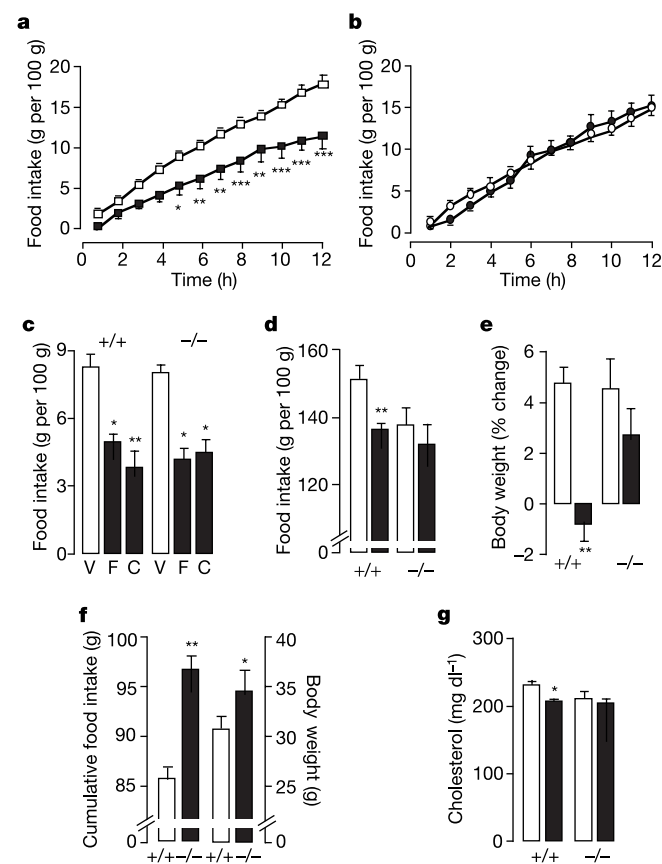


Figure 2 PPAR- α mediates the appetite-suppressing and weight-reducing effects of OEA. **a–c**, OEA reduces feeding in wild-type but not PPAR- α -/- mice. **a**, **b**, Time course of the effects of OEA (filled squares; 10 mg per kg i.p.) and vehicle (open squares) on total food intake in wild-type ($n = 6–11$; **a**) and PPAR- α -/- ($n = 5–8$; **b**) mice. **c**, Effects of vehicle (V), d-fenfluramine (F; 4 mg per kg i.p.) and cholecystikinin-8 (C; 25 μ g per kg i.p.) on food intake in wild-type and PPAR- α -/- mice ($n = 4–9$). **d**, **e**, OEA reduces food intake and body weight gain in obese wild-type but not PPAR- α -/- mice. Shown are the effects of OEA (filled bars; 5 mg per kg i.p.) and vehicle (open bars) on food intake (**d**) and body weight gain (**e**). **f**, PPAR- α -/- mice fed with a high-fat diet develop hyperphagia. Food intake and body weight in wild-type and PPAR- α -/- mice before OEA treatment ($n = 4–9$). **g**, Effects of OEA (filled bars; 5 mg per kg i.p.) and vehicle (open bars) on serum cholesterol levels in obese wild-type and PPAR- α -/- mice. * $P < 0.05$ and ** $P < 0.001$ ($n = 6–8$) by one-way ANOVA followed by Dunnett's test or Student's t -test (**c–g**).

engaged PPAR- α with EC₅₀ values of $1.4 \pm 0.1 \mu\text{M}$ and $150 \pm 20 \text{ nM}$, respectively ($n = 5$). Saturation binding studies showed that [³H]OEA associates with the purified LBD of mouse and human PPAR- α (Fig. 1c), with dissociation constants (K_d) of $37.4 \pm 0.1 \text{ nM}$ (mouse, $n = 10$) and $43.3 \pm 1.6 \text{ nM}$ (human, $n = 8$). Binding competition experiments yielded a half-maximal inhibitory concentration (IC₅₀) for OEA of $120.0 \pm 10.7 \text{ nM}$ ($n = 10$; Fig. 1d). The results indicate that OEA is a high-affinity agonist of PPAR- α .

To test whether PPAR- α activation contributes to the appetite-suppressing properties of OEA, we used mice deficient in PPAR- α (PPAR- $\alpha^{-/-}$ mice)^{13,14}. OEA reduced feeding in wild-type mice, but not in PPAR- $\alpha^{-/-}$ mice (Fig. 2a, b), although the two strains showed comparable tissue concentrations of OEA (Supplementary Table 1) and similar responses to other appetite suppressants, such as D-fenfluramine and cholecystokinin-8 (Fig. 2c). In addition to its short-term effects on feeding, OEA produces in rats a sustained inhibition of body weight gain¹. To determine the role of PPAR- α in this response, we fed wild-type and PPAR- $\alpha^{-/-}$ mice with a high-fat chow and, once they had become obese, treated them daily with OEA for 4 weeks. OEA reduced total food intake and suppressed body weight gain in wild-type mice, whereas it had no effect in PPAR- $\alpha^{-/-}$ mice (Fig. 2d, e). Notably, whereas mutant mice receiving a standard chow maintain normal body weight^{13,14}, those fed with a high-fat diet ate more and gained more weight than did wild-type mice (Fig. 2f). Thus, PPAR- α expression may not be necessary only for the satiety-inducing and weight-reducing actions of OEA, but also for the maintenance of feeding homeostasis.

Does OEA modulate feeding through direct activation of PPAR- α ? To address this question, we used three synthetic PPAR- α agonists of different potency^{11,12}: Wy-14643 and GW7674, which engage PPAR- α *in vitro* with potencies comparable to that of OEA, inhibited food intake in C57BL/6J mice (Fig. 3a), whereas the weaker agonist clofibrate did not (25–100 mg per kg body weight; data not shown). Meal pattern analyses showed that the effects of Wy-14643 and GW7647 were due to prolonged eating

latency, rather than to changes in meal size or post-meal interval (Fig. 3b). This response is identical to that elicited by OEA (Fig. 3b) and is indicative of a satiety-inducing action².

OEA is thought to produce satiety by activating peripheral sensory fibres¹. Accordingly, in rats in which these fibres had been removed either by cutting the vagus nerve below the diaphragm or by treatment with the neurotoxin capsaicin¹, OEA had no effect on food intake (Fig. 3c and data not shown). These procedures also prevented the effect of Wy-14643 (Fig. 3d, e, and Supplementary Table 2), whereas they did not affect that of D-fenfluramine, which reduces appetite through a central mechanism (Fig. 3c).

The close correspondence between the effects of OEA and those of synthetic PPAR- α agonists is reinforced by two findings. First, potent agonists of PPAR- β/δ (GW501516)¹⁵ and PPAR- γ (ciglitazone)¹⁶ did not affect feeding in C57BL/6J mice (Fig. 3f); and second, PPAR- $\alpha^{-/-}$ mice did not respond to Wy-14643 (Fig. 3g, h). We cannot exclude the possibility that PPAR- β/δ , which weakly recognizes OEA *in vitro* (Fig. 1a), may contribute to some OEA responses; however, the fact that the PPAR- β/δ agonist GW501516 does not affect food intake implies that the role of PPAR- β/δ in OEA signalling, if any, is distinct from that of PPAR- α .

To test further the ability of OEA to activate PPAR- α , we examined its effects on the expression of PPAR- α target genes. We focused first on the small intestine, which is a likely site of action of OEA¹ and contains large amounts of PPAR- α ^{17,18}. In the jejunum of C57BL/6J mice, OEA increased the expression of three genes regulated by PPAR- α (Fig. 4a–c): those encoding PPAR- α , fatty acid translocase (FAT/CD36) and fatty acid transport protein 1 (FATP1)^{17,19}. A similar effect was observed in the liver (Fig. 4e–g) and the duodenum, but not in the ileum (Supplementary Figs 1 and 2). By contrast, OEA did not affect the expression of three additional genes that are not controlled by PPAR- α (Fig. 4d).

Underscoring the role of PPAR- α in these responses, we found that Wy-14643 mimicked the effects of OEA (Fig. 4a–g) and that neither OEA nor Wy-14643 changed gene expression in PPAR- $\alpha^{-/-}$ mice (Fig. 4a–c, e–g). In addition to stimulating transcription, PPAR- α also represses genes such as that encoding inducible nitric

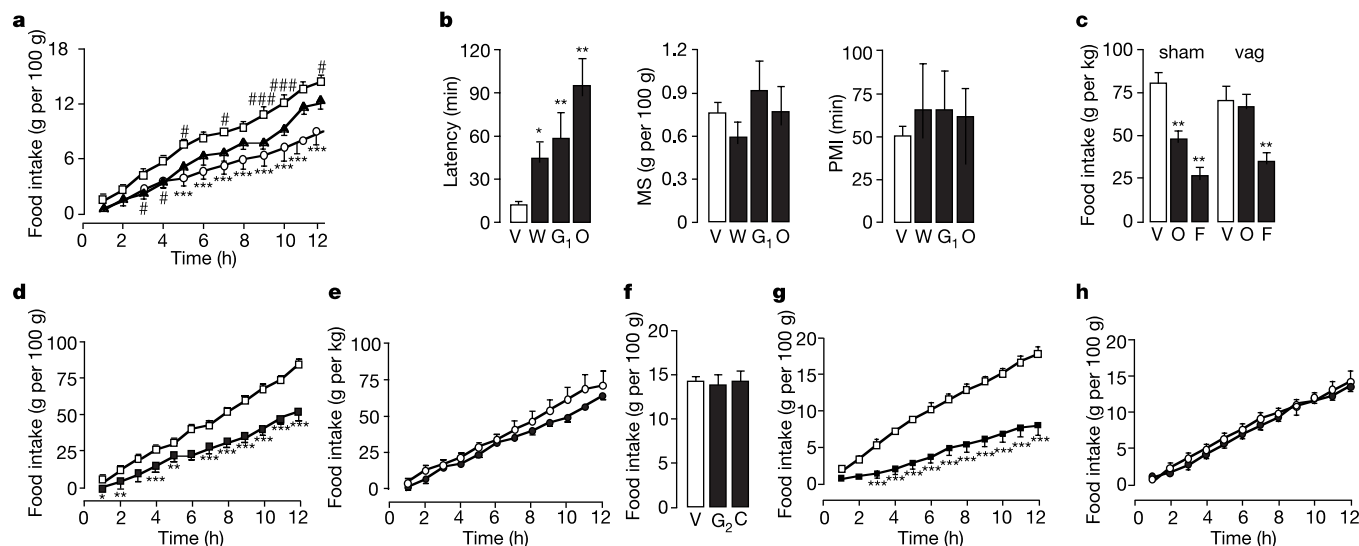


Figure 3 PPAR- α agonists mimic the appetite-suppressing effects of OEA. **a**, Effects of vehicle (open squares), Wy-14643 (filled triangles; 40 mg per kg i.p.) and GW7647 (open circles; 20 mg per kg i.p.) on food intake in C57BL/6J mice (vehicle, $n = 40$; drugs, $n = 4-7$). **b**, Effects of vehicle (V), Wy-14643 (W; 40 mg per kg i.p.), GW7647 (G₁; 20 mg per kg i.p.) and OEA (O; 10 mg per kg i.p.) on feeding latency, first meal size (MS) and first post-meal interval (PMI) in C57BL/6J mice (vehicle, $n = 40$; drugs, $n = 4-7$). **c**, Effects of vehicle (V), OEA (O; 10 mg per kg i.p.) and D-fenfluramine (F; 3 mg per kg subcutaneous) on food intake in control (sham) and vagotomized (vag) rats. **d**, **e**, Time

course of the effects of vehicle (open symbols) and Wy-14643 (filled symbols; 40 mg per kg i.p.) on food intake in control (**d**) and vagotomized (**e**) rats. **f**, Lack of effect of the PPAR- β/δ agonist GW501516 (G₂; 5 mg per kg i.p.) and PPAR- γ agonist ciglitazone (C; 15 mg per kg i.p.) on food intake in C57BL/6J mice ($n = 4-6$). **g**, **h**, Time course of the effects of vehicle (open symbols) and Wy-14643 (filled symbols; 40 mg per kg i.p.) on food intake in wild-type (**g**) and PPAR- $\alpha^{-/-}$ (**h**) mice ($n = 7-11$). * $P < 0.05$, ** $P < 0.001$ and *** $P < 0.0001$ by one-way ANOVA followed by Dunnett's test or Student's *t*-test with Bonferroni's correction.

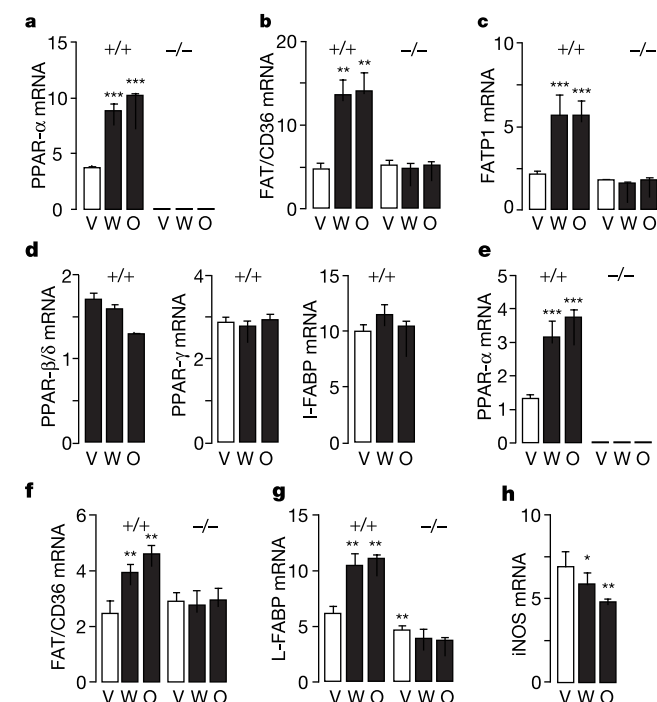


Figure 4 OEA regulates gene expression through PPAR-α activation. **a–d**, Effects of vehicle (V), Wy-14643 (W; 30 mg per kg i.p.) and OEA (O; 10 mg per kg i.p.) on messenger RNA levels of PPAR-α (**a**), FAT/CD36 (**b**), FATP1 (**c**) and PPAR-β/δ, PPAR-γ and I-FABP (**d**) in the jejunum of wild-type and PPAR-α^{-/-} mice (*n* = 5). **e–g**, Effects of vehicle (V), Wy-14643 (W; 30 mg per kg i.p.) and OEA (O; 10 mg per kg i.p.) on mRNA levels of PPAR-α (**e**), FAT/CD36 (**f**) and L-FABP (**g**) in the liver of wild-type and PPAR-α^{-/-} mice (*n* = 5). **h**, Transrepression of iNOS expression by OEA (O; 10 mg per kg i.p.) and Wy-14643 (W; 30 mg per kg i.p.) in the jejunum of C57BL/6J mice (*n* = 5). mRNA levels are expressed in arbitrary units. **P* < 0.05 and ***P* < 0.001 by one-way ANOVA followed by Dunnett's test.

oxide synthase (iNOS)²⁰. Accordingly, in the jejunum of C57BL/6J mice, OEA and Wy-14643 decreased iNOS expression (Fig. 4h). These results indicate that OEA closely mimics the genomic actions of PPAR-α agonists.

If OEA enhances expression of PPAR-α target genes, it should also reproduce the corrective effects of PPAR-α agonists on serum lipid levels^{21–23}. Consistent with this prediction, in genetically obese Zucker (*fafa*) rats, a 2-week regimen of OEA administration decreased serum cholesterol and triglyceride along with food intake and body weight gain (Supplementary Fig. 3). In addition, 4 weeks of OEA treatment lowered cholesterol in obese wild-type mice²⁴, but not in PPAR-α^{-/-} mice (Fig. 2g). We conclude that long-term OEA administration causes metabolic changes similar to those produced by PPAR-α agonists, which are abrogated by PPAR-α deletion.

The functional similarity between OEA and PPAR-α agonists suggests that OEA is a natural ligand for PPAR-α. This possibility is supported by the concerted regulation of OEA synthesis and expression of PPAR-α and iNOS. In the small intestine of C57BL/6J mice, OEA concentrations were significantly lower at night (01:30), when the mice feed, than during the day (16:30), when they are satiated and resting (Fig. 5a, b). The expression of PPAR-α paralleled OEA concentrations, whereas that of its transrepression target iNOS showed the opposite pattern (Fig. 5c, d). Notably, daytime intestinal concentrations of OEA were in the range needed to activate PPAR-α fully *in vitro* (~300 nM), suggesting that they may be adequate to engage the receptor *in vivo*.

In conclusion, our findings indicate that OEA shows all of the defining characteristics of an endogenous PPAR-α agonist: first, it binds with high affinity to PPAR-α; second, it mimics the actions of

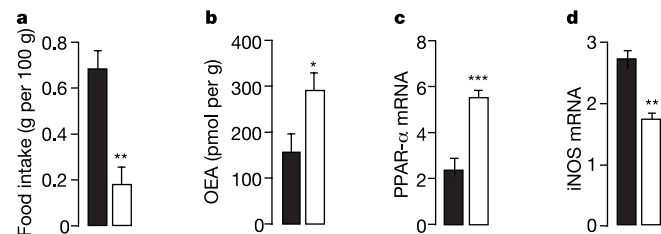


Figure 5 Circadian regulation of intestinal OEA synthesis and PPAR-α expression. Food intake (**a**), OEA content (**b**), and PPAR-α (**c**) and iNOS (**d**) mRNA levels at 01:30 (filled bars) and 16:30 (open bars) in free-feeding C57BL/6J mice maintained on a 12-h/12-h dark/light cycle.

synthetic agonists in a PPAR-α-dependent manner; and last, under appropriate conditions, it reaches tissue concentrations that are sufficient to activate PPAR-α. The results also suggest that PPAR-α activation not only mediates OEA-induced weight stabilization, which may be expected from the metabolic roles of this receptor^{6–8}, but is also responsible for OEA-induced satiety, a behavioural role that was not previously attributed to PPAR-α. The molecular mechanism underlying this response is still undefined, but one possibility is that it may involve the transcriptional regulation of intestinal NO production. Intestinal epithelial cells contain the enzyme iNOS and generate significant amounts of its product, NO, which may act as a peripheral appetite-stimulating signal^{25,26}. The ability of OEA to reduce iNOS expression through PPAR-α suggests that downregulation of NO signalling may contribute to the persistent satiety-inducing actions of this lipid mediator. □

Methods

Chemicals

We synthesized GW7647 (ref. 12) and GW501516 (ref. 27) by modifications to the published procedures (see Supplementary Information) and synthesized OEA as described²⁸. All other chemicals were purchased from Sigma or Tocris.

Animals

We purchased male C57BL/6J, PPAR-α^{-/-} (129S4/SvJae-*Ppara*^{tm1Gonz}) and wild-type 129S1/SvImJ mice (Jackson Laboratories), male Zucker rats (Charles River) and vagotomized Sprague-Dawley rats (Zivic Laboratories). Animals were maintained on a 12-h/12-h light/dark cycle and had free access to water and RMH 2500 chow (Prolab).

Transactivation assays

We generated plasmids containing the LBD of human PPAR-α (nucleotides 499–1,407), PPAR-β/δ (412–1,323), PPAR-γ (610–1,518) and RXR (403–1,389) fused to the DNA-binding domain of yeast GAL4 under control of the human cytomegalovirus promoter and to a neomycin resistance gene to provide stable selection with 0.2 mg ml⁻¹ G418 (Calbiochem). We cultured HeLa cells in DMEM medium supplemented with fetal bovine serum (10%), transfected them with 3 μl of Eugene 6 (Roche) containing 1 μg of pFR-luc plasmid (Stratagene), and replaced the media after 18 h with DMEM containing 0.1 mg ml⁻¹ hygromycin (Calbiochem). After 4 weeks, we isolated surviving clones, analysed them by luciferase assay, and selected for transfection a cell line that showed the highest luciferase activity. Cells were maintained in DMEM containing hygromycin and G418.

For transactivation assays, we seeded cells in six-well plates and incubated them for 7 h in DMEM containing hygromycin and G418, plus appropriate concentrations of test compounds. We used a dual-luciferase reporter assay system (Promega) and an MIX Microtiter plate luminometer (Dynex) to determine luciferase activity in cell lysates.

Binding assays

We constructed a pGEX-α plasmid containing the gene encoding glutathione S-transferase (GST) fused to the LBD of human PPAR-α by subcloning the LBD fragment into pGEX-4T (Amersham Biosciences) digested with *EcoRI*. *Escherichia coli* strain BL21 (Novagen) was cultured in 2XYT medium and induced for overexpression by the addition of 0.1 mM isopropyl-1-thio-β-D-galactopyranoside (IPTG). The IPTG-induced culture was grown at 30 °C for an additional 6 h and cells were collected by centrifugation. The GST-LBD fusion protein was purified from the cell pellet by using glutathione-Sepharose beads (Amersham Biosciences), as recommended by the manufacturer.

For binding assays, the GST-LBD fusion protein (50 nM) was incubated at 25 °C for 2 h in buffer containing 50 mM HEPES, pH 7.0, 50 mM KCl, 5 mM EDTA, 10 mM dithiothreitol (DTT), 0.1% bovine serum albumin and 15 Ci mmol⁻¹ [³H]OEA (ARC). Free and bound [³H]OEA were separated on a Sephadex G-25 spin column (Amersham Biosciences) and radioactivity in the bound [³H]OEA fraction was measured. Nonspecific binding was determined in the presence of non-radioactive OEA (50 μM).

RNA isolation and complementary DNA synthesis

We stored tissues in RNALater (Ambion), extracted total RNA with TRIzol (Invitrogen) and quantified it with Ribogreen (Molecular Probes). We synthesized cDNA with SuperscriptII RNase H reverse transcriptase (Invitrogen).

Polymerase chain reaction

Reverse transcription of 2 µg of total RNA was carried out with 0.2 µg of Oligo(dT)₁₂₋₁₈ primer for 50 min at 42 °C, and real-time quantitative polymerase chain reaction was done with an ABI PRISM 7700 sequence detection system (Applied Biosystems). We designed primer/probe sets with Primer Express software (Applied Biosystems) and gene sequences available from GenBank database. Primers and fluorogenic probes were synthesized by TIB Molbiol.

The primer/probe sequences for the mouse genes were as follows. For PPAR-α: forward (F), 5'-CTTCCCAAAGCTCCTTCAAAAA-3'; reverse (R), 5'-CTGCGCATGCTCCGTG-3'; probe (P), 5'-TGGTGGACCTTCGCGAGCTGG-3'. For PPAR-β/δ: F, 5'-GATGACA GTGACCTGGCGCT-3'; R, 5'-AGGCTTGGCCGGTCTC-3'; P, 5'-TTCATCGCGGCCAT CATCTCTGTG-3'. For PPAR-γ: F, 5'-AGTGGAGACCGCCAGG-3'; R, 5'-GCAGCAGG TTGCTCTGGATG-3'; P, 5'-TTGCTGAACGTGAAGCCCATCGA-3'. For FAT1/CD36: F, 5'-CGGCGATGAGAAAGCAGAA-3'; R, 5'-CAACCAGGCCAGGAGC-3'; P, 5'-TGTTTCAAGAACCAAGTGACCGGGAAATAA-3'. For FATP1: F, 5'-GCACAGCAGG TACTACCGCA-3'; R, 5'-GGCGGCAGCATGC-3'; P, 5'-TGCTGCCTTTGGCCACCA TTCCTA-3'. For L-FABP: F, 5'-TCACCATCACCTATGGACCCA-3'; R, 5'-TCCAGTTCCG CACTCCTCCC-3'; P, 5'-AGTGGTCCGCAATGAGTTCACCCCTG-3'. For GAPDH: F, 5'-TCACTGGCATGGCCTTCC-3'; R, 5'-GGCGGCACGTACATCC-3'; P, 5'-TTCCTA CCCCAAATGTGTCCGCTCG-3'. RNA levels were normalized by using glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as an internal standard, and measurements were made as described²⁹.

Feeding experiments

For acute experiments, we administered drugs or vehicles (saline for CCK-8 and D-fenfluramine, dimethylsulphoxide (DMSO)/saline (70/30) for all other agents; 4 ml per kg intraperitoneally (i.p.)) at 17:00–17:30 to free-feeding animals habituated to the experimental setting. Feeding was monitored for 12 h by an automated system (SciPro)².

For chronic experiments, we fed male wild-type and PPAR-α^{-/-} mice with a D12492 high-fat diet (60 kcal % fat; Research Diets) for 7 weeks. Body mass indices³⁰ were 0.36 ± 0.01 g cm⁻² for wild-type (*n* = 13) and 0.41 ± 0.01 g cm⁻² for PPAR-α^{-/-} (*n* = 15) mice. We divided the mice into four groups (*n* = 7–8 each) and treated them for 4 weeks with vehicle (saline/polyethylene glycol/Tween 80 (90/5/5); 1 ml per kg) or OEA (5 mg per kg i.p., once daily). In a separate experiment, we treated obese Zucker rats for 2 weeks with vehicle or OEA (5 mg per kg i.p., once daily), while maintaining them on RMH 2500 regular chow (Prolab). We measured food intake and body weight daily.

For capsaicin deafferentation, male Wistar rats were treated with capsaicin as described¹. The rats were deprived of food for 24 h and given Wy-14643 (2–40 mg per kg i.p.) or vehicle. Food pellets and spillage were measured manually 0.5–4 h after drug injection.

Biochemical analyses

We measured OEA levels by high-performance liquid chromatography coupled with mass spectrometry³, and serum lipids with a commercial kit (Wako).

Statistical analyses

Results are expressed as the mean ± s.e.m of *n* separate experiments. The significance of differences between groups was evaluated by one-way analysis of variance (ANOVA) followed by a Dunnett's test for multiple comparisons, or Student's *t*-test with or without Bonferroni's correction. Analyses were done with GraphPad Prism software (GraphPad).

Received 19 April; accepted 15 July 2003; doi:10.1038/nature01921.

- Rodríguez de Fonseca, F. *et al.* An anorexic lipid mediator regulated by feeding. *Nature* **414**, 209–212 (2001).
- Gaetani, S., Oveisi, F. & Piomelli, D. Modulation of meal pattern in the rat by the anorexic lipid mediator oleoylethanolamide. *Neuropsychopharmacology* **28**, 1311–1316 (2003).
- Göttlicher, M., Widmark, E., Li, Q. & Gustafsson, J. A. Fatty acids activate a chimera of the clofibrate acid-activated receptor and the glucocorticoid receptor. *Proc. Natl Acad. Sci. USA* **89**, 4653–4657 (1992).
- Kliwer, S. A. *et al.* Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors α and γ. *Proc. Natl Acad. Sci. USA* **94**, 4318–4323 (1997).
- Forman, B. M., Chen, J. & Evans, R. M. Hypolipidemic drugs, polyunsaturated fatty acids, and eicosanoids are ligands for peroxisome proliferator-activated receptors α and δ. *Proc. Natl Acad. Sci. USA* **94**, 4312–4317 (1997).
- Desvergne, B. & Wahli, W. Peroxisome proliferator-activated receptors: nuclear control of metabolism. *Endocr. Rev.* **20**, 649–688 (1999).
- Chawla, A., Repa, J. J., Evans, R. M. & Mangelsdorf, D. J. Nuclear receptors and lipid physiology: opening the X-files. *Science* **294**, 1866–1870 (2001).
- Berger, J. & Moller, D. E. The mechanisms of action of PPARs. *Annu. Rev. Med.* **53**, 409–435 (2002).
- Lazennec, G., Canaple, L., Saugy, D. & Wahli, W. Activation of peroxisome proliferator-activated receptors (PPARs) by their ligands and protein kinase A activators. *Mol. Endocrinol.* **14**, 1962–1975 (2000).
- Devane, W. A. *et al.* Isolation and structure of a brain constituent that binds to the cannabinoid receptor. *Science* **258**, 1946–1949 (1992).
- Willson, T. M., Brown, P. J., Sternbach, D. D. & Henke, B. R. The PPARs: from orphan receptors to drug discovery. *J. Med. Chem.* **43**, 527–550 (2000).
- Brown, P. J. *et al.* Chemical compounds as selective activators of PPAR-α. *PCT Int. Appl.* **32** (2000).
- Lee, S. S. *et al.* Targeted disruption of the α isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators. *Mol. Cell. Biol.* **15**, 3012–3022 (1995).

- Butler, A. A. & Cone, R. D. Knockout models resulting in the development of obesity. *Trends Genet.* **17**, S50–S54 (2001).
- Oliver, W. R. Jr *et al.* A selective peroxisome proliferator-activated receptor δ agonist promotes reverse cholesterol transport. *Proc. Natl Acad. Sci. USA* **98**, 5306–5311 (2001).
- Chang, A. Y., Wyse, B. M., Gilchrist, B. J., Peterson, T. & Diani, A. R. Ciglitazone, a new hypoglycemic agent. I. Studies in *ob/ob* and *db/db* mice, diabetic Chinese hamsters, and normal and streptozotocin-diabetic rats. *Diabetes* **32**, 830–838 (1983).
- Martin, G., Schoonjans, K., Lefebvre, A. M., Staels, B. & Auwerx, J. Coordinate regulation of the expression of the fatty acid transport protein and acyl-CoA synthetase genes by PPARα and PPARγ activators. *J. Biol. Chem.* **272**, 28210–28217 (1997).
- Escher, P. *et al.* Rat PPARs: quantitative analysis in adult rat tissues and regulation in fasting and refeeding. *Endocrinology* **142**, 4195–4202 (2001).
- Motojima, K., Passilly, P., Peters, J. M., González, F. J. & Latruffe, N. Expression of putative fatty acid transporter genes are regulated by peroxisome proliferator-activated receptor α and γ activators in a tissue- and inducer-specific manner. *J. Biol. Chem.* **273**, 16710–16714 (1998).
- Colville-Nash, P. R., Qureshi, S. S. & Willoughby, D. A. Inhibition of inducible nitric oxide synthase by peroxisome proliferator-activated receptor agonist: correlation of induction of heme oxygenase 1. *J. Immunol.* **161**, 978–984 (1998).
- Best, J. D. & Jenkins, A. J. Novel agents for managing dyslipidaemia. *Expert Opin. Invest. Drugs* **10**, 1901–1911 (2001).
- Cleary, M. P., Kasiske, B., O'Donnell, M. P. & Keane, W. F. Effect of long-term clofibrate acid treatment on serum and tissue lipid and cholesterol levels in obese Zucker rats. *Atherosclerosis* **66**, 107–112 (1987).
- Chaput, E., Saladin, R., Silvestre, M. & Edgar, A. D. Fenofibrate and rosiglitazone lower serum triglycerides with opposing effects on body weight. *Biochem. Biophys. Res. Commun.* **271**, 445–450 (2000).
- Akiyama, T. E. *et al.* Peroxisome proliferator-activated receptor-α regulates lipid homeostasis, but is not associated with obesity: studies with congenic mouse lines. *J. Biol. Chem.* **276**, 39088–39093 (2001).
- Sticker-krongrad, A., Beck, B. & Bulet, C. Nitric oxide mediates hyperphagia of obese Zucker rats: relation to specific changes in the microstructure of feeding behavior. *Life Sci.* **58**, PL9–PL15 (1996).
- Janero, D. R. Nutritional aspects of nitric oxide: human health implication and therapeutic opportunities. *Nutrition* **17**, 896–903 (2001).
- Chao, E. Y.-H. *et al.* Thiazoles and oxazole derivatives and their pharmaceutical use. *PCT Int. Appl.* **83** (2001).
- Giuffrida, A., Rodríguez de Fonseca, F. & Piomelli, D. Quantification of bioactive acylethanolamides in rat plasma by electrospray mass spectrometry. *Anal. Biochem.* **280**, 87–93 (2000).
- Schmittgen, T. D. *et al.* Quantitative reverse transcription-polymerase chain reaction to study mRNA decay: comparison of endpoint and real-time methods. *Anal. Biochem.* **285**, 194–204 (2000).
- Gregoire, F. M. *et al.* Diet-induced obesity and hepatic gene expression alterations in C57BL/6J and ICAM-1-deficient mice. *Am. J. Physiol. Endocrinol. Metab.* **282**, E703–E713 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Stein for comments; L. Giron, N. Heyrani, N. Izadi and K. Nguyen for help with experiments; M. Guzmán for critically reading the manuscript; and F. Valiño for synthesizing fatty acid ethanolamides. This research was supported by grants (to D.P.) from the National Institute on Drug Abuse. Further support came from the Fondo de Investigación Sanitaria.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to D.P. (piomelli@uci.edu).

Proton-sensing G-protein-coupled receptors

Marie-Gabrielle Ludwig¹, Miroslava Vanek¹, Danilo Guerini¹, Jürg A. Gasser¹, Carol E. Jones², Uwe Junker¹, Hans Hofstetter¹, Romain M. Wolf¹ & Klaus Seuwen¹

¹Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland

²Novartis Horsham Research Centre, Wimbleshurst Rd, Horsham, West Sussex RH12 5AB, UK

Blood pH is maintained in a narrow range around pH 7.4 mainly through regulation of respiration and renal acid extrusion^{1,2}. The molecular mechanisms involved in pH homeostasis are not completely understood. Here we show that ovarian cancer G-protein-coupled receptor 1 (OGR1), previously described as a receptor for sphingosylphosphorylcholine³, acts as a proton-sensing receptor stimulating inositol phosphate formation. The receptor is inactive at pH 7.8, and fully activated at pH 6.8—site-

directed mutagenesis shows that histidines at the extracellular surface are involved in pH sensing. We find that GPR4, a close relative of OGR1, also responds to pH changes, but elicits cyclic AMP formation. It is known that the skeleton participates in pH homeostasis as a buffering organ, and that osteoblasts respond to pH changes in the physiological range⁴, but the pH-sensing mechanism operating in these cells was hitherto not known. We detect expression of OGR1 in osteosarcoma cells and primary human osteoblast precursors, and show that these cells exhibit strong pH-dependent inositol phosphate formation. Immunohistochemistry on rat tissue sections confirms the presence of OGR1 in osteoblasts and osteocytes. We propose that OGR1 and GPR4 are proton-sensing receptors involved in pH homeostasis.

Following transient or stable expression of OGR1 we noted a robust and apparently ligand-independent activation of phosphoinositide turnover in transfected cells. This effect persisted in assay buffer devoid of either calcium, magnesium, phosphate or sulphate, excluding the possibility that OGR1 was activated by any of these constituents (data not shown). However, variations of buffer pH strongly affected inositol phosphate (IP) formation, as illustrated in Fig. 1a. CCL39 hamster fibroblasts stably expressing OGR1 were incubated at different pH, and accumulation of IPs measured at 37 °C in the absence or presence of the inositol monophosphatase inhibitor lithium^{5,6}. At pH 7.6, IP formation in the presence of lithium was close to background, but lower pH resulted in a significant accumulation of IPs, which was linear over 30 min and maximal at pH 6.8. The pH of our standard assay buffer is 7.4, explaining the basal activity observed in previous experiments. In Fig. 1b, activation of OGR1 by protons is plotted as a standard concentration-response curve, covering a wider range of extracellular pH. Half-maximal activation of the receptor occurred at pH 7.48 ± 0.04 (mean $\pm$ s.e.m.; $N = 12$). Activation appeared highly co-operative, with a Hill coefficient of >2 . Under acidic conditions (pH <6.5), IP formation declined again, reflecting the pH-dependence of the phosphoinositide signalling system in CCL39 cells (see also Fig. 1c, thrombin). Sphingosylphosphorylcholine (SPC), the bioactive lipid reported to activate OGR1³, did not stimulate the receptor in our experiments, and did not affect pH-dependent stimulation of IP formation. However, the molecule inhibited forskolin-stimulated adenylyl cyclase in other cell systems (data not shown), suggesting it acts on a receptor distinct from OGR1.

Importantly, untransfected cells (Fig. 1c) or cells expressing other receptors (data not shown) did not show pH-dependent IP formation. Similarly, the response to other receptor agonists was not positively modulated in the range of pH 7–8, as demonstrated for thrombin (Fig. 1c), which activates endogenous receptors in CCL39 cells⁷. Pertussis toxin (PTX) did not inhibit IP formation measured in the presence of lithium at pH 7 (Fig. 1d), indicating that OGR1 acts through G_q ⁸. The response to thrombin, which is additive to the signal elicited by OGR1 in these cells, was partially inhibited by PTX both at pH 7 and at pH 7.6, as expected⁷.

Data reported in Fig. 1 show that activation of OGR1 at neutral or slightly acidic pH is strong, comparable to activation of other G-protein-coupled receptors (GPCRs) by their cognate ligands. Strong pH-dependent IP formation was also observed in stably and transiently transfected HEK293 cells (see below; half-maximal activation at pH 7.44 ± 0.06 (mean $\pm$ s.e.m.; $N = 5$)). Some pH-activated channels are sensitive to temperature⁹, and we therefore tested whether OGR1 was affected by temperature in the range of 35–42 °C. This was not the case; IP formation rate was only minimally affected in this range (data not shown). For cells expressing ectopic OGR1 we could also demonstrate pH-dependent stimulation of intracellular calcium transients, reporter gene transcription, and GDP–GTP exchange in membrane preparations (see Supplementary Information).

We expected that histidines should play an important role in pH sensing. Indeed, inspection of a three-dimensional (3D) model of

OGR1 (Fig. 2) indicated a cluster of histidines at the extracellular surface, placed on top of helices I, IV and VII, and in extracellular loops 1 and 2. All are conserved in the mouse, rat, and bovine sequence (not shown, accession nos. XM_138218, XM_234483 (Refseq) and U88367 (GenBank), respectively). Specifically, the model predicts a direct hydrogen-bond interaction between H20 and H269 in the unprotonated state, thus linking helix I and helix VII (Fig. 2b). A second interaction is possible between H17 and H84, linking the receptor amino terminus to extracellular loop 1.

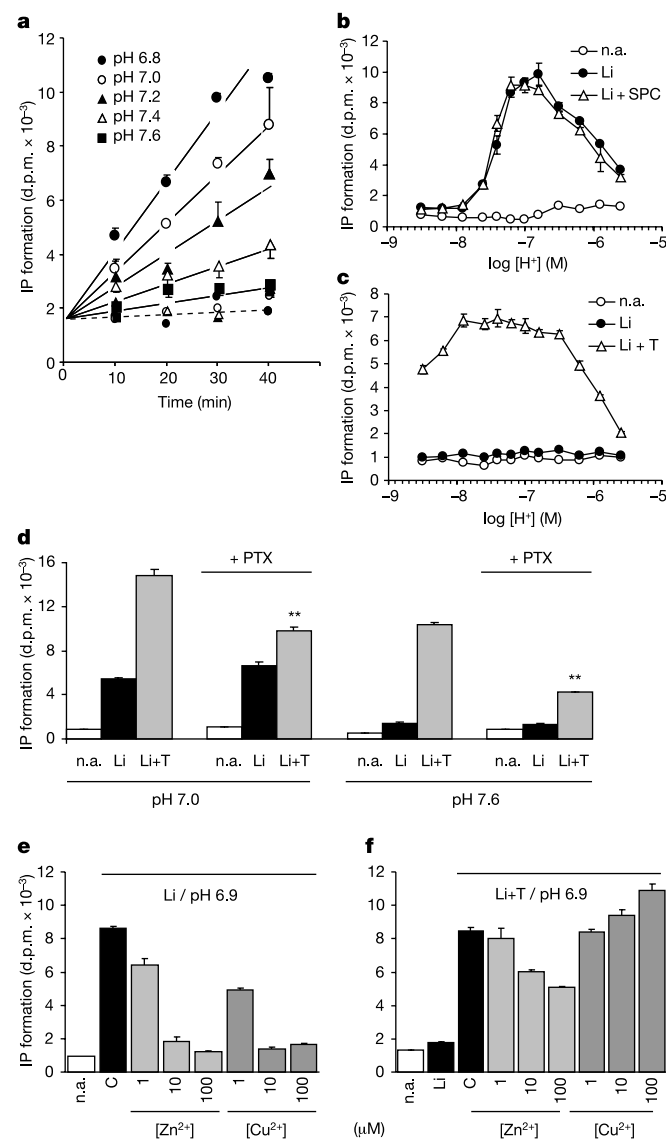


Figure 1 IP formation experiments. **a, b, d, e**, CCL39 cells stably expressing OGR1; **c, f**, untransfected control cells. **a**, Time course. Cells were incubated at indicated pH and lithium was added to some wells after 10 min ($t = 0$). Plates were removed from the experiment after further 10, 20, 30, 40 min and total IPs determined. Smaller symbols at bottom of figure (dashed line) indicate signal in the absence of lithium. **b**, Concentration-response curve for protons. Cells were incubated in HEM-PSS at indicated pH without further addition (no addition; n.a.) or in the presence of lithium (Li) or Li plus SPC (1 μ M). **c**, Effect of pH on thrombin (T; 1 U ml⁻¹) response in control cells; buffer as in **b, d**. **d**, Effect of PTX. Cells were pretreated or not with PTX and incubated at pH 7.0 or 7.6 without further addition (n.a.) or in the presence of Li or Li + T (1 U ml⁻¹). Asterisks indicate significant difference ($P < 0.01$; t -test) compared to the corresponding signals in the absence of PTX. **e, f**, Inhibition of IP formation by Cu²⁺ and Zn²⁺ ions. CCL39 cells expressing OGR1 (**e**) or control cells (**f**) were incubated in PSS containing the indicated concentrations of ZnCl₂ or CuCl₂. In **f**, IP formation was stimulated by T (3 U ml⁻¹).

However, given the flexibility of the extracellular loops, this interaction appears less likely to form, and initial constraints were required in our model to make it fall in place. Additional histidines were found at other locations in the receptor, but their mutual interaction or electrostatic interactions with other residues seemed less obvious. We set out to mutate all potentially critical histidines individually to phenylalanine, and to measure pH-dependent receptor activation in transient transfection experiments. Recombinant proteins were found to be expressed on the cell membrane at similar levels, as verified by immunocytochemistry using the carboxy-terminal myc tag. Expression of specific mutants was further controlled by confocal microscopy (Fig. 3c and data not shown).

IP formation in HEK293 cells transiently transfected with OGR1 (Fig. 3a) was more resistant to acidic pH than shown above for stable expression in CCL39, and this fact was instrumental for the study of receptor mutants. It is probably due to receptor overexpression following transient transfection, as the cells' endogenous response to carbachol was inhibited at acidic pH (Fig. 3b) similar to the thrombin response in CCL39 (Fig. 1c). As shown in Fig. 3d, mutation of histidines 89, 159, 175 remained without major effect on receptor function. However, histidines 17, 20, 84, 269 were each required for normal pH sensitivity, as expected. In addition H169

turned out to be important. Mutation of these amino acids resulted in receptors that failed to stimulate IP formation at pH 6.8; but on exposure to more acidic pH, activity could be restored. Response curves appeared shifted to the right, that is, sensitivity towards the ligand was reduced. In no case could an increased pH-independent basal activity be observed. A straightforward explanation for the involvement of H169 in proton sensing is difficult on the basis of our model, given its position in extracellular loop 2, the geometry of which can not be predicted at this stage. Mutation of H245 was not tolerated, leading to a severe, albeit not total, loss of function. According to our model, this amino acid is located in helix VI facing the lipid environment, and may be required for overall structural integrity. It is conserved in all receptors related to OGR1 (GPR4, TDAG8, G2A; alignment not shown). Double mutations of paired histidines (17,84; 20,269) resulted in activation profiles similar to that observed with corresponding single mutations. Surprisingly, mutation of all four paired histidines did not fully eliminate function. Almost total receptor inactivation occurred with mutation of all five histidines initially identified as participating in pH sensing. In order to generate a receptor with a minimal number of histidines and still functional in the pH range of 7–7.8, we tested a H89,159,175F construct. The encoded protein indeed still functioned as the wild type (Fig. 3d, bottom right).

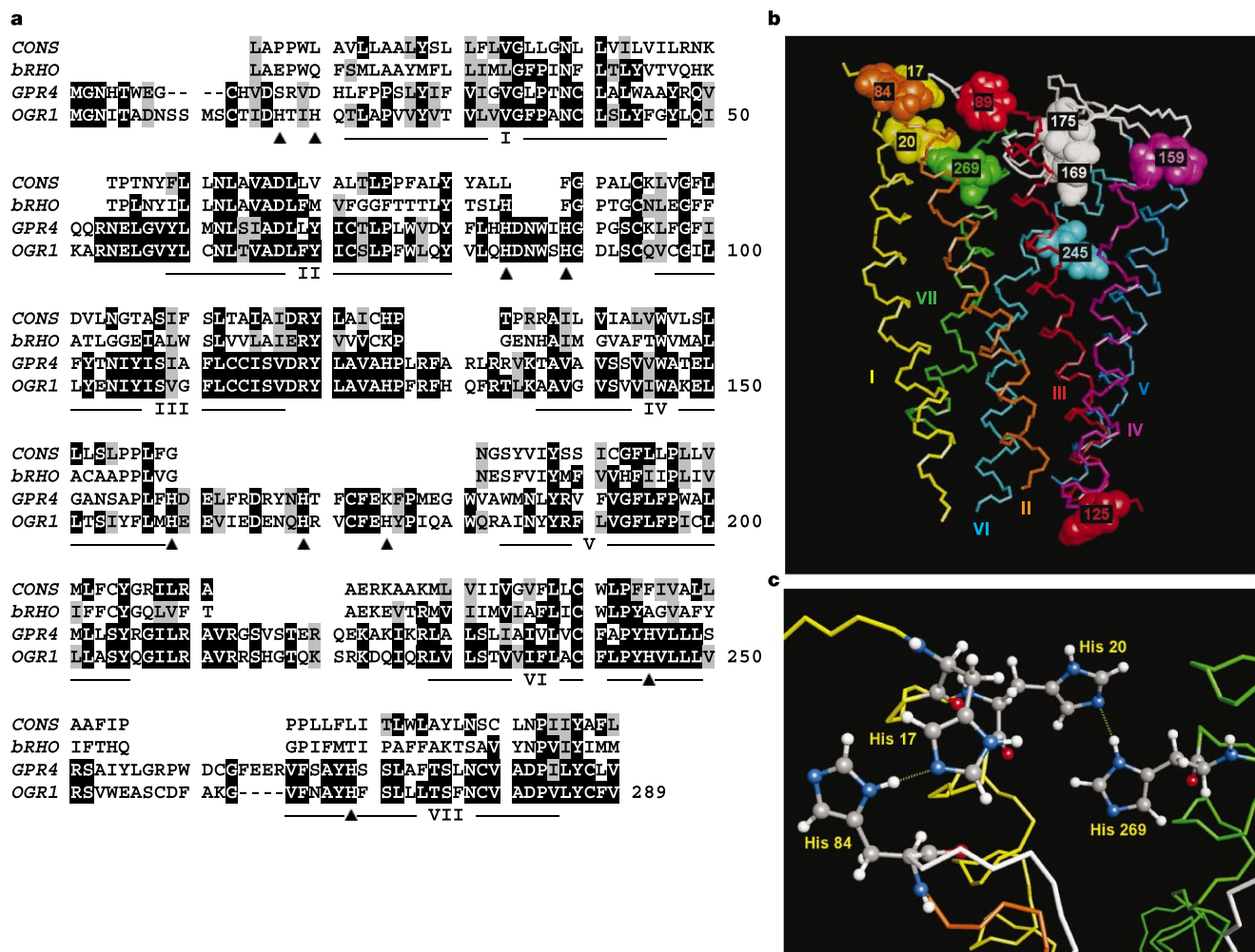


Figure 2 Sequence alignment and 3D model for OGR1. **a**, Sequence alignment of OGR1 and GPR4, without C-terminal intracellular domain. Numbering according to OGR1. Sequence fragments shown for bovine RHO (brHO) and for the rhodopsin family consensus (<http://www.gpcr.org/7tm/seq>) indicate the regions used for homology modelling. Amino acids fully conserved between OGR1 and the other sequences are

highlighted in black, conservative substitutions in grey. The transmembrane domains (I–VII) are underlined and as proposed by SwissProt for OGR1 (Q15743). Histidines mutated in this study are marked by a triangle. **b**, Backbone of the OGR1 model with all histidine residues highlighted. **c**, Close up view of the predicted interactions between H17 and H84, and between H20 and H269.

Pairs of histidines are able to co-ordinate Zn^{2+} and Cu^{2+} atoms, and this fact has been used to study GPCR structure-function^{10,11}. On the basis of our model, we expected Zn^{2+} and Cu^{2+} to inhibit pH-dependent receptor activation, by stabilizing the unprotonated state of histidine pairs. Indeed, micromolar concentrations of both ions inhibited OGR1-dependent IP formation stimulated at pH 6.9 (Fig. 1e). In control CCL39 cells, Cu^{2+} remained without effect on the thrombin response, and Zn^{2+} ions led to a partial inhibition (Fig. 1f).

Taken together, our receptor model and the experimental data

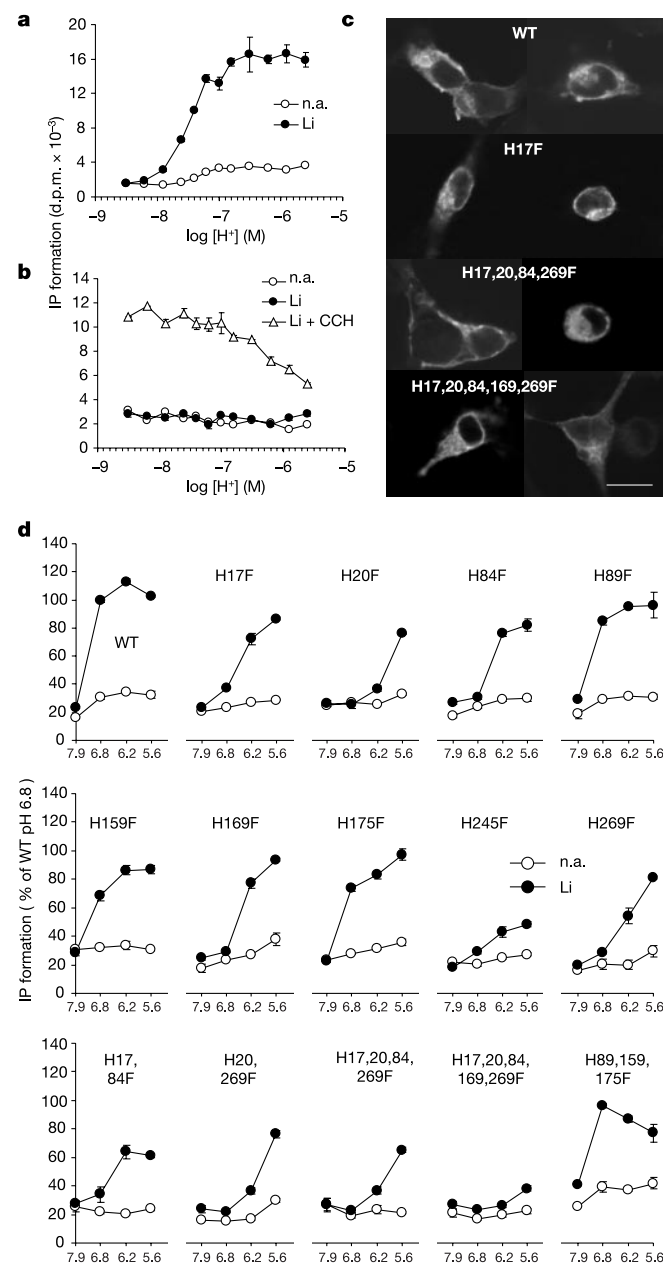


Figure 3 pH-dependent activation of wild-type and mutated OGR1 receptor constructs following transient transfection into HEK293 cells. IP formation was measured in HEM-PSS at indicated pH in the absence (n.a.) or presence of lithium (Li). **a**, Wild-type receptor, full concentration-response curve; **b**, endogenous response to carbachol (CCH, 1 mM) in untransfected cells. **c**, Confocal microscopy images showing membrane location of receptor proteins (myc-tag). Scale bar, 10 μm . **d**, IP formation responses for mutated receptor constructs. Data were normalized to the signal observed at pH 6.8 with the wild-type receptor transfected in parallel (100%).

suggest that at slightly alkaline pH OGR1 is stabilized in an inactive state by hydrogen bonding involving histidines. Protonation leads to loss of bonding and probably repulsion between residues, allowing the receptor to adopt an active conformation.

GPR4¹² is closely related to OGR1, with all histidines important for pH-sensing conserved (Fig. 2a), albeit the histidines equivalent to H17 and H20 appear slightly displaced in the primary sequence. Expression studies showed that this receptor indeed responds to pH as OGR1, but its primary mode of coupling is G_s , that is, activation of cAMP formation (Fig. 4a). SPC, also reported to activate GPR4¹², did not modulate pH-dependent stimulation and did not activate cAMP formation on its own. Half-maximal activation of cAMP formation by GPR4 in transiently transfected HEK293 cells occurred at pH 7.55 $\pm$ 0.02 (mean $\pm$ s.e.m.; $N = 6$). Untransfected cells (Fig. 4b) or cells expressing unrelated receptors (data not shown) did not show any pH-dependent stimulation of cAMP formation. Similarly, neither was the response of other receptors to agonists positively modulated in the range of pH 7.8–7.0, as demonstrated here for the β -adrenergic receptor agonist isoproterenol (Fig. 4b).

Expression profiling by reverse transcriptase-mediated polymerase chain reaction (RT-PCR) revealed the presence of messenger RNA for OGR1 (Fig. 5a), but not GPR4 (not shown), in MG63 human osteosarcoma cells, and indeed we found that these cells responded to neutral or acidic pH with IP formation (Fig. 5b). The signal was comparable in amplitude to that elicited by bradykinin (Fig. 5d, e), which activates endogenous receptors in MG63 cells¹³. Half-maximal activation was observed at pH 7.46 $\pm$ 0.01 (mean $\pm$ s.e.m.; $N = 5$). IP formation was insensitive to PTX pretreatment (Fig. 5d) and inhibited by micromolar concentrations of copper ions (Fig. 5e). Similar results were obtained for primary human preosteoblasts isolated from trabecular bone (Fig. 5a, c, and data not shown). Half-maximal activation in primary cells occurred at pH 7.41 $\pm$ 0.02 (mean $\pm$ s.e.m.; $N = 3$).

Immunohistochemistry (IHC) on rat bone sections revealed OGR1 protein in active osteoblasts, lining cells on the bone surface,

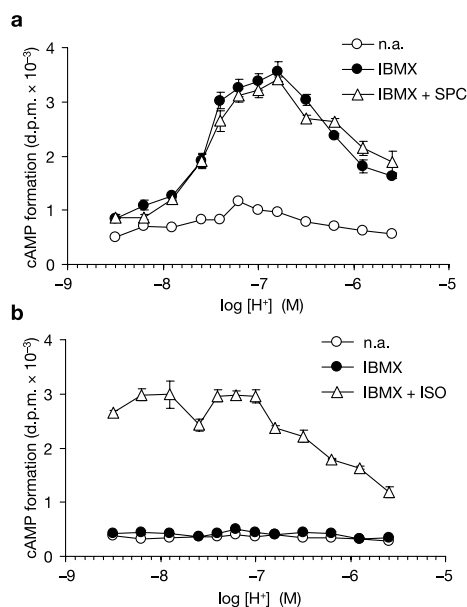


Figure 4 cAMP formation experiments. **a**, HEK293 cells transiently transfected with GPR4; **b**, untransfected control cells. Data are shown as means $\pm$ s.e.m. for triplicate determinations. **a**, Concentration-response curves for protons. Cells were incubated in HEM-PSS at indicated pH without further addition (n.a.) or in the presence of IBMX or IBMX plus SPC (1 μM). **b**, Effect of pH on the isoproterenol (ISO, 10 μM) response in control cells.

and matrix-embedded osteocytes (Fig. 5f). The receptor was also specifically expressed in chondrocytes of hypertrophic cartilage, epithelial cells of the lung, intestine, and renal tubuli, as well as in skeletal myocytes and hepatocytes (to be published elsewhere).

Bone stores large amounts of bicarbonate, and participates in pH homeostasis¹⁴. Acidosis was shown to increase osteoclastic bone resorption and to suppress osteoblast function, and these events may be linked to osteoblastic PGE₂ production^{4,15,16} downstream of

phosphoinositide breakdown¹⁷. The pH-sensing mechanism operating in these cells remained unknown. Our data suggest that OGR1 represents an osteoblastic pH sensor regulating cell-mediated responses to acidosis in bone.

In summary, we have demonstrated the pH-sensing properties of two GPCRs, OGR1 and GPR4. Our findings do not exclude the existence of other modulators of receptor function. Until now, only specific ion channels were known to respond to changes in extracellular pH in the physiological range, and their role in nociception is clearly established^{18,19}—but these molecules appear to operate mainly in the nervous system. Both OGR1 and GPR4 show a more widespread tissue distribution^{3,12,20}. Besides bone metabolism, they may be involved in the regulation of respiration¹ and cardiovascular functions²¹, as well as other physiological functions². Weak-to-moderate extracellular acidification is also observed in all pathological states linked to deterioration of the blood supply, like inflammation and ischaemia^{22,23}. Expression of OGR1 and GPR4 in immune cells may also be relevant in this context. □

Methods

Cell culture and transfections

CCL39 hamster fibroblasts, HEK293 human embryonic kidney cells, and MG63 human osteosarcoma cells were grown in a 1:1 mixture of bicarbonate-buffered DMEM and Ham's F12 medium supplemented with 10% fetal calf serum and antibiotics, in CO₂ atmosphere at pH 7.4. Primary cultures of human trabecular bone-derived preosteoblasts were established as described²⁴. Expression vectors for human OGR1 and GPR4 (NM_003485 and NM_005282, respectively) were prepared using pcDNA3.1(+)/myc-His (Invitrogen). Site-directed mutagenesis was carried out using the Quick Change kit from Stratagene. For stable transfection, vectors were linearized with *PvuII*. Stable and transient transfections were carried out using the Effectene reagent (Qiagen). Stable cell populations expressing receptors were isolated following selection with antibiotic G418 (400 µg ml⁻¹).

Buffers and pH

Experiments were carried out in a physiological salt solution (PSS) containing 130 mM NaCl, 0.9 mM NaH₂PO₄, 5.4 mM KCl, 0.8 mM MgSO₄, 1.0 mM CaCl₂, 25 mM glucose. This solution was buffered either with HEPES alone (20 mM) or HEPES/EPPS/MES (8 mM each; HEM-PSS), to cover a wider pH range. HEPES-buffered PSS was used in all experiments unless HEM-PSS is specifically mentioned. HEPES is 4-(2-hydroxyethyl) piperazine-1-ethanesulphonic acid, EPPS is N'-(2-hydroxyethyl)piperazine-N'-3-propanesulphonic acid, MES is 2-(N-morpholino)ethanesulphonic acid. The pH of all solutions was adjusted using a carefully calibrated pH meter (Metrohm). All data in this report are referenced to pH at room temperature. To obtain pH at 37°C, 0.15 pH units should be subtracted for HEPES buffers in the range of pH 6.8–7.8 according to our calibration experiments.

IP formation assay

Confluent cell cultures grown in 24-well plates were labelled with myo[³H]inositol (100 MBq ml⁻¹; ART/Anawa Trading) for 24 h in serum-free DMEM medium. Where indicated, cells were pretreated with PTX (100 ng ml⁻¹; Sigma) during the last 4 h of inositol loading. Cells were then incubated at 37°C in PSS with indicated buffer. Lithium (20 mM) was added to block inositol monophosphatase activity, leading to accumulation of IP₁^{5,6}. Where indicated, bovine thrombin, SPC, carbachol or bradykinin (Sigma) was added 1 min before lithium addition. Unless otherwise stated, incubation was continued for 20 min. Cells were then extracted with ice-cold formic acid and total IPs separated from free inositol using batch column chromatography²⁵. Data are shown as means ± s.e.m. for triplicate determinations.

cAMP formation assay

Confluent cell cultures grown in 24-well plates were labelled with [3H]adenine (100 MBq ml⁻¹; Amersham) for 4 h in serum-free DMEM medium. Cells were then incubated for 15 min at 37°C in HEM-PSS containing the phosphodiesterase inhibitor isobutylmethylxanthine (IBMX, 1 mM) where indicated, to allow accumulation of cAMP. Following extraction with ice-cold trichloroacetic acid, cAMP was separated from free adenine and ATP using batch column chromatography²⁶.

RT-PCR expression profiling

Total RNA was prepared from cell cultures using the acid phenol method. RNA was DNase-treated and reverse-transcribed using Superscript II (Invitrogen). Parallel reactions were performed without enzyme, to allow control for genomic DNA contamination. PCR reactions for OGR1 and glyceraldehyde-3-phosphate dehydrogenase (GPDH) were set up with Expand High Fidelity Taq (Roche) using the following temperature cycling protocol: 30 s denaturation at 94°C, 45 s annealing at 65°C (OGR1) or 55°C (GPDH), 50 s extension at 72°C; 36 cycles for OGR1, 30 cycles for GPDH. GPDH was measured as internal standard for mRNA quantity. For OGR1, PCR reaction products were cloned and verified by sequencing. The following primers were used: OGR1 forward: 5'-CTGAGCCCATGAGGAGTGTG-3', reverse: 5'-GGTAGGACGCCAGCAGCAGG-3'; GPDH forward: 5'-TTAGCACCCCTGGCCAAGG-3', reverse: 5'-CTTACTCTCTGGAGGCCATG-3'.

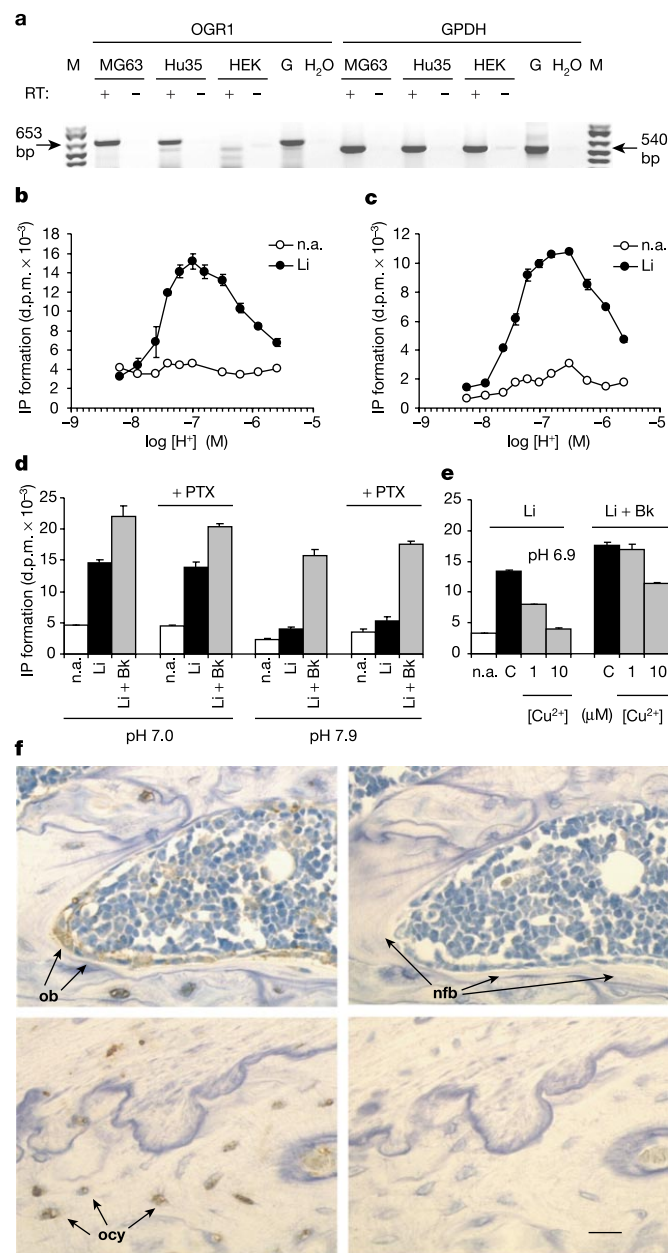


Figure 5 OGR1 in bone. **a**, RT-PCR detection of OGR1 and GPDH mRNA in MG63, primary human preosteoblasts (Hu35) and HEK293 cells. M, molecular weight ladder; RT, reverse transcriptase; G, genomic DNA control. **b**, **c**, Concentration-response curves for protons. MG63 (**b**) or Hu35 cells (**c**) were incubated in HEM-PSS at indicated pH without further addition (n.a.) or in the presence of lithium (Li). **d**, Effect of PTX. MG63 cells were pretreated or not with PTX and incubated at pH 7.0 or 7.9 without further addition (n.a.) or in the presence of Li or Li plus bradykinin (Bk; 1 µM). **e**, Inhibition of IP formation by Cu²⁺ ions in MG63 cells. **f**, Detection by IHC of OGR1 in sections of rat cancellous bone (top panel) and cortical bone (lower panel). Right panels show adjacent sections not exposed to the primary antibody. ob, osteoblasts; ocy, osteocytes; nfb, newly formed bone. Scale bar, 20 µm.

Immunolocalization studies

Recombinant receptors were detected using a FITC-labelled anti-myc antibody (Zymed, no. 132511). Experiments on bone tissue were performed on 4 µm sections of paraffin-embedded organs collected from 6-month-old female Wistar rats. IHC was performed using an affinity-purified rabbit polyclonal antibody developed and tested for IHC by Lifespan Biosciences Inc., Seattle, directed against the peptide epitope CFVSETTHRDRLRLRG, which is identical on human and rat OGR1. Staining was revealed using the ABC peroxidase staining kit (Santa Cruz Biotechnology). Slides were counterstained with Gill's haematoxylin.

Receptor model

Given the acceptable homology between OGR1 and other rhodopsin-like GPCRs, we used the published bovine rhodopsin (RHO) structure²⁷ as a template. The OGR1 primary sequence was aligned to bovine RHO as shown in Fig. 2a, and for the indicated amino acid positions the side chains present in bovine RHO were substituted by the corresponding side chains of OGR1, using the SCWRL program²⁸. The resulting structure was refined applying molecular mechanics (energy refinement) and dynamics, using the parm96.dat parameters of the AMBER force field²⁹ with the distance-dependent dielectric function of $\epsilon = 1^*$ and the sander_classic module of the AMBER6 package. In all computations, the C α atoms were constrained in motion by a moderate harmonic potential. The extracellular loops and all amino-acid side chains were left free to move within the potential of the force field. A disulphide bond was introduced between residues CYS94 and CYS172. Intracellular loops were not included.

Received 7 May; accepted 8 July 2003; doi:10.1038/nature01905.

- Feldman, J. L., Mitchell, G. S. & Nattie, E. E. Breathing: rhythmicity, plasticity, chemosensitivity. *Annu. Rev. Neurosci.* **26**, 239–266 (2003).
- Swenson, E. R. Metabolic acidosis. *Respir. Care* **46**, 342–353 (2001).
- Xu, Y. et al. Sphingosylphosphorylcholine is a ligand for ovarian cancer G-protein-coupled receptor 1. *Nature Cell Biol.* **2**, 261–267 (2000).
- Krieger, N. S., Parker, W. R., Alexander, K. M. & Bushinsky, D. A. Prostaglandins regulate acid-induced cell-mediated bone resorption. *Am. J. Physiol. Renal Physiol.* **279**, F1077–F1082 (2000).
- Berridge, M. J., Downes, C. P. & Hanley, M. R. Lithium amplifies agonist-dependent phosphatidylinositol responses in brain and salivary glands. *Biochem. J.* **206**, 587–595 (1982).
- Berridge, M. J. & Irvine, R. F. Inositol phosphates and cell signalling. *Nature* **341**, 197–205 (1989).
- Paris, S. & Pouyssegur, J. Pertussis toxin inhibits thrombin-induced activation of phosphoinositide hydrolysis and Na⁺/H⁺ exchange in hamster fibroblasts. *EMBO J.* **5**, 55–60 (1986).
- Simon, M. I., Strathmann, M. P. & Gautam, N. Diversity of G proteins in signal transduction. *Science* **252**, 802–808 (1991).
- Smith, G. D. et al. TRPV3 is a temperature-sensitive vanilloid receptor-like protein. *Nature* **418**, 186–190 (2002).
- Elling, C. E., Nielsen, S. M. & Schwartz, T. W. Conversion of antagonist-binding site to metal-ion site in the tachykinin NK-1 receptor. *Nature* **374**, 74–77 (1995).
- Sheikh, S. P., Zyzya, T. A., Lichtarge, O., Sakmar, T. P. & Bourne, H. R. Rhodopsin activation blocked by metal-ion-binding sites linking transmembrane helices C and F. *Nature* **383**, 347–350 (1996).
- Zhu, K. et al. Sphingosylphosphorylcholine and lysophosphatidylcholine are ligands for the G protein-coupled receptor GPR4. *J. Biol. Chem.* **276**, 41325–41335 (2001).
- Brechter, A. B. & Lerner, U. H. Characterization of bradykinin receptors in a human osteoblastic cell line. *Regul. Pept.* **103**, 39–51 (2002).
- Bushinsky, D. A. Acid-base imbalance and the skeleton. *Eur. J. Nutr.* **40**, 238–244 (2001).
- Goldhaber, P. & Rabadjia, L. H⁺ stimulation of cell-mediated bone resorption in tissue culture. *Am. J. Physiol.* **253**, E90–E98 (1987).
- Meghji, S., Morrison, M. S., Henderson, B. & Arnett, T. R. pH dependence of bone resorption: Mouse calvarial osteoclasts are activated by acidosis. *Am. J. Physiol. Endocrinol. Metab.* **280**, E112–E119 (2001).
- Gijon, M. A. & Leslie, C. C. Regulation of arachidonic acid release and cytosolic phospholipase A2 activation. *J. Leucoc. Biol.* **65**, 330–336 (1999).
- Waldmann, R. et al. H⁺-gated cation channels. *Ann. NY Acad. Sci.* **868**, 67–76 (1999).
- Caterina, M. J. et al. Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science* **288**, 306–313 (2000).
- An, S., Tsai, C. & Goetzl, E. J. Cloning, sequencing and tissue distribution of two related G protein-coupled receptor candidates expressed prominently in human lung tissue. *FEBS Lett.* **375**, 121–124 (1995).
- Ishizaka, H., Gudi, S. R., Frangos, J. A. & Kuo, L. Coronary arteriolar dilation to acidosis: role of ATP-sensitive potassium channels and pertussis toxin-sensitive G proteins. *Circulation* **99**, 558–563 (1999).
- Reeh, P. W. & Steen, K. H. Tissue acidosis in nociception and pain. *Prog. Brain Res.* **113**, 143–151 (1996).
- Heming, T. A. et al. Effects of extracellular pH on tumour necrosis factor- α production by resident alveolar macrophages. *Clin. Sci. (Lond.)* **101**, 267–274 (2001).
- Sottile, V., Halleux, C., Bassilana, F., Keller, H. & Seuwen, K. Stem cell characteristics of human trabecular bone-derived cells. *Bone* **30**, 699–704 (2002).
- Seuwen, K., Lagarde, A. & Pouyssegur, J. Deregulation of hamster fibroblast proliferation by mutated ras oncogenes is not mediated by constitutive activation of phosphoinositide-specific phospholipase C. *EMBO J.* **7**, 161–168 (1988).
- Salomon, Y. Adenylate cyclase assay. *Adv. Cycl. Nucleot. Res.* **10**, 35–55 (1979).
- Palczewski, K. et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**, 739–745 (2000).
- Bower, M. J., Cohen, F. E. & Dunbrack, R. L. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.* **267**, 1268–1282 (1997).
- Cornell, W. D. et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank B. Wilmering-Wetter and H. Ankin for assistance in cell culture, IP formation and intracellular calcium assays, R. Brearley and M. Schaublin for support in molecular biology, and A. Rebmann for IHC analysis. We also thank R. Bouhelal and I. Vranesic for advice concerning the fluorescence imaging plate reader, and J. Mosbacher for help with confocal microscopy.

Authors' contributions The experiments shown in Figs 1, 3–5 were carried out by M.G.L., M.V., D.G. and K.S. J.A.G. performed the IHC on tissue sections. C.E.J. cloned GPR4 and performed expression profiling experiments. U.J. and H.H. participated in the cloning of OGR1, expression profiling, and experiments with lipid agonists. R.W. carried out receptor modelling.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.S. (Klaus.seuwen@pharma.novartis.com).

Structure of the catalytic domain of human phosphodiesterase 5 with bound drug molecules

Byung-Je Sung^{1*}, Kwang Yeon Hwang^{1*}, Young Ho Jeon^{1*}, Jae Il Lee^{1*}, Yong-Seok Heo¹, Jin Hwan Kim¹, Jinho Moon¹, Jung Min Yoon¹, Young-Lan Hyun¹, Eunmi Kim¹, Sung Jin Eum¹, Sam-Yong Park², Jie-Oh Lee³, Tae Gyu Lee¹, Seonggu Ro¹ & Joong Myung Cho¹

¹The Division of Drug Discovery, CrystalGenomics, Inc., Daedeok Biocommunity, Jeonmin-dong, Yuseong-gu, Daejeon, 305-390, South Korea

²Protein Design Laboratory, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

³Department of Chemistry, Korea Advanced Institute of Science and Technology, 373-1 Kusong-dong, Yuseong-gu, Daejeon 305-701, South Korea

* These authors contributed equally to this work

Phosphodiesterases (PDEs) are a superfamily of enzymes that degrade the intracellular second messengers cyclic AMP and cyclic GMP^{1–3}. As essential regulators of cyclic nucleotide signalling with diverse physiological functions, PDEs are drug targets for the treatment of various diseases, including heart failure, depression, asthma, inflammation and erectile dysfunction^{4–7}. Of the 12 PDE gene families, cGMP-specific PDE5 carries out the principal cGMP-hydrolysing activity in human corpus cavernosum tissue. It is well known as the target of sildenafil citrate (Viagra) and other similar drugs for the treatment of erectile dysfunction. Despite the pressing need to develop selective PDE inhibitors as therapeutic drugs, only the cAMP-specific PDE4 structures are currently available^{8,9}. Here we present the three-dimensional structures of the catalytic domain (residues 537–860) of human PDE5 complexed with the three drug molecules sildenafil, tadalafil (Cialis) and vardenafil (Levitra). These structures will provide opportunities to design potent and selective PDE inhibitors with improved pharmacological profiles.

PDE5 is composed of an amino-terminal regulatory domain (GAF A and B) and a carboxy-terminal metal-binding catalytic domain¹⁰. PDE5A is the only subtype of PDE5 reported, and there are four isoforms (PDE5A1–4). The catalytic domains of these splicing variants are exactly the same. The PDE5–sildenafil complex structure was determined at 2.3 Å by the multi-wavelength anomalous dispersion (MAD) method using Se-Met protein crystals (Fig. 1; see also Supplementary Information). The structures of tadalafil and vardenafil complexes were determined at 2.8 Å and 2.5 Å, respectively. Sildenafil binds to PDE5 with a dissociation constant (K_d) range of 0.5–6.6 nM, and to PDE6, the most similar

enzyme to PDE5, about 9–11 times more weakly¹¹. Tadalafil binds to PDE5 with a K_d range of about 0.9–6.7 nM, 200–700 times more tightly than it binds PDE6 (ref. 11).

The catalytic domain of PDE5 has three helical subdomains, an N-terminal cyclin-fold region, a linker region and a C-terminal helical bundle (Fig. 1). Residues from Ile 665 to Leu 675 in the linker domain are disordered and are not visible in electron density maps. The overall topology is quite similar to that of the homologous enzyme PDE4, although PDE5 has only 23% sequence identity with PDE4 in the catalytic region. The N-terminal domain adopts the cyclin-fold topology, previously observed in cyclin A and TFIIB^{12,13}. It is composed of a bundle of five α -helices ($\alpha 1$, $\alpha 3$, $\alpha 5$, $\alpha 6$ and $\alpha 8$) with three short 3_{10} -helices ($\alpha 2$, $\alpha 4$ and $\alpha 7$). The core helix $\alpha 3$ is flanked on one side by helices $\alpha 5$, $\alpha 6$ and $\alpha 8$, which form an interface with the linker domain ($\alpha 9$ and $\alpha 10$) and the C-terminal domain ($\alpha 12$ and $\alpha 13$). The C-terminal helical bundle is formed by five long α -helices ($\alpha 11$, $\alpha 12$, $\alpha 14$, $\alpha 17$ and $\alpha 18$) and three short interconnecting α -helices ($\alpha 13$, $\alpha 15$ and $\alpha 16$). The linker domain is composed of long anti-parallel α -helices ($\alpha 9$ and $\alpha 10$), which are separated by the disordered region.

The active site of PDE5 is located at the centre of the C-terminal helical bundle domain. The substrate pocket is approximately 10 Å deep, with a narrow opening and a wide inner space, giving a

total volume of about 330 Å³. It is composed of four subsites: a metal-binding site (M site), core pocket (Q pocket), hydrophobic pocket (H pocket) and lid region (L region) (Supplementary Fig. 2). Overall, the M site and Q pocket are similar to those of PDE4, but the H pocket and the L region show significant structural differences compared with PDE4 (Fig. 3).

The M site contains both a zinc ion and a second metal (presumably magnesium) ion, and is surrounded by helices $\alpha 6$, $\alpha 8$, $\alpha 9$, $\alpha 10$ and $\alpha 12$. The zinc ion is bound to the side chains of His 617 (bond length 2.1 Å), Asp 654 (2.1 Å), Asp 764 (2.2 Å), His 653 (2.1 Å) and two water molecules (W1 2.2 Å and W2 2.8 Å). Asp 654 and water molecule W2 coordinate both metal ions. The second metal ion is coordinated by Asp 654 (2.0 Å) and five water molecules (2.0, 2.0, 2.1, 2.0 and 2.6 Å). Three of these water molecules form hydrogen bonds with residues His 657, Asp 682 and His 685. Residues at this site are highly conserved in all PDE family members (Supplementary Fig. 1). None of the three inhibitors described interacts directly with the M site.

The Q pocket accommodates the pyrazolopyrimidinone group of sildenafil, strongly suggesting that the chemically similar guanidine group of cGMP also binds at this region. This pocket is lined by Gln 817, Phe 820, Val 782 and Tyr 612, which are highly conserved in all PDEs. The amide moiety of the pyrazolopyrimidinone group

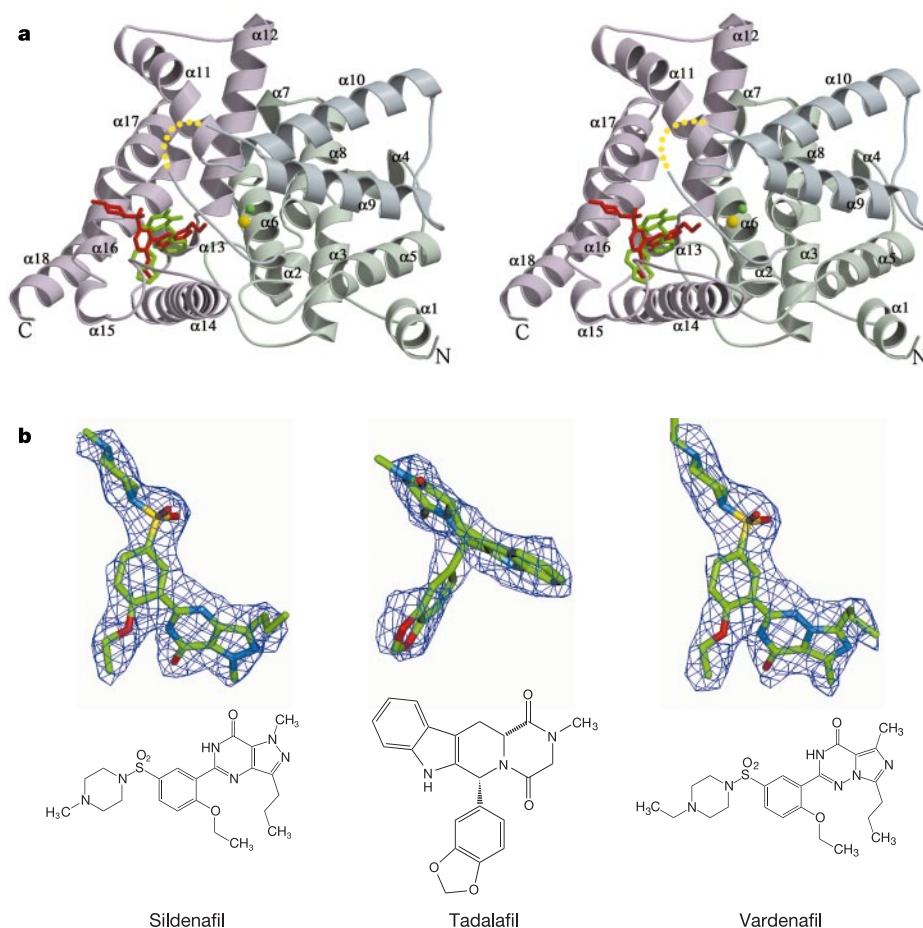


Figure 1 Overview of the PDE5 complex structures. **a**, Stereo ribbon diagram of the human PDE5 structure. The catalytic domain of the PDE5 molecule can be divided into three subdomains: an N-terminal cyclin-fold domain (residues 537–678, light green), a linker helical domain (residues 679–725, light blue) and C-terminal helical bundle domain (residues 726–860, violet). The bound sildenafil and tadalafil molecules are overlapped and shown as stick models (red and green, respectively). The zinc (the bigger CPK model)

and magnesium ions are represented as yellow and green spheres, respectively. An 11-residue loop in the linker domain not visible in the structure is indicated by yellow dots. **b**, The sigma-A-weighted $2F_o - F_c$ electron density maps covering the sildenafil, tadalafil and vardenafil molecules at a level of 1.0σ , together with their chemical structures.

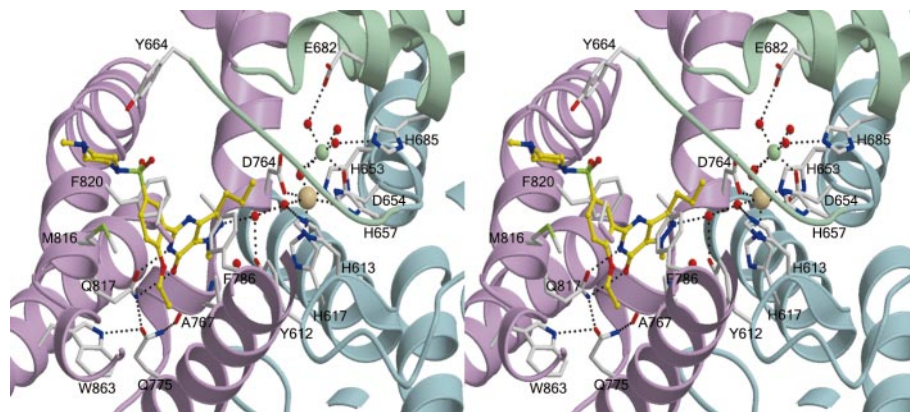


Figure 2 Stereo view of the active site of the PDE5–sildenafil complex. Sildenafil is shown as a stick model with carbon atoms coloured yellow. Metal- and inhibitor-binding residues of PDE5 are shown as stick models with carbon atoms coloured white. The zinc (the bigger CPK model) and magnesium ions are shown in orange and green, respectively. The amide

moiety of the pyrazolopyrimidinone group of sildenafil forms a bidentate hydrogen bond with the γ -amide group of Gln 817, which is well ordered by a hydrogen bond relay involving Gln 817 to Gln 775, Gln 775 to Ala 767 and Gln 775 to Trp 853. The model orientation is the same as in Fig. 1a.

forms a bidentate hydrogen bond with the γ -amide group of Gln 817. This residue is completely conserved in all the PDE families, and an equivalent hydrogen bond was observed in previously reported PDE4–ligand complex structures^{8,9}. Moreover, its side chain is well ordered by a hydrogen bond relay involving Gln 817 to Gln 775, Gln 775 to Ala 767 and Gln 775 to Trp 853, facilitating the binding of sildenafil. This relay also accounts for the cGMP substrate specificity⁸, as it holds the Gln 817 side chain in a position to hydrogen bond with a donor and acceptor 1-NH and 6-CO groups on the purine base of cGMP (Supplementary Fig. 3). The hydrogen-bonding character at the 1- and 6-positions is reversed in cAMP. In PDE4B, Tyr 403 forms a hydrogen bond with Gln 443, which hydrogen bonds with the donor and acceptor 1-N and 6-NH₂ groups on the cAMP substrate. The ability of a glutamine side chain to accept and donate hydrogen bonds there-

fore seems to have an important role in discriminating between cGMP and cAMP.

N₂ of the pyrazole moiety of sildenafil forms a hydrogen bond with a water molecule that is in turn hydrogen-bonded to both the side chain of Tyr 612 and another water molecule coordinating to Zn ion (Fig. 2). Site-directed mutagenesis of Tyr 612 causes a marked increase in Michaelis constant (K_m) value for cGMP¹⁴. The pyrazole ring also forms hydrophobic interactions with the side chains of Val 782, Leu 785, Tyr 612 and Phe 820, including a face-to-face π – π interaction with the phenyl ring of Phe 820.

The ethoxyphenyl group of sildenafil fits into the hydrophobic H pocket formed by Phe 786, Ala 783, Leu 804 and Val 782. Leu 804 is replaced by Met in PDE4, and superposition of the crystal structures of PDE4 and PDE5 shows that the side chain of the Met residue would sterically hinder the phenyl ring. This could

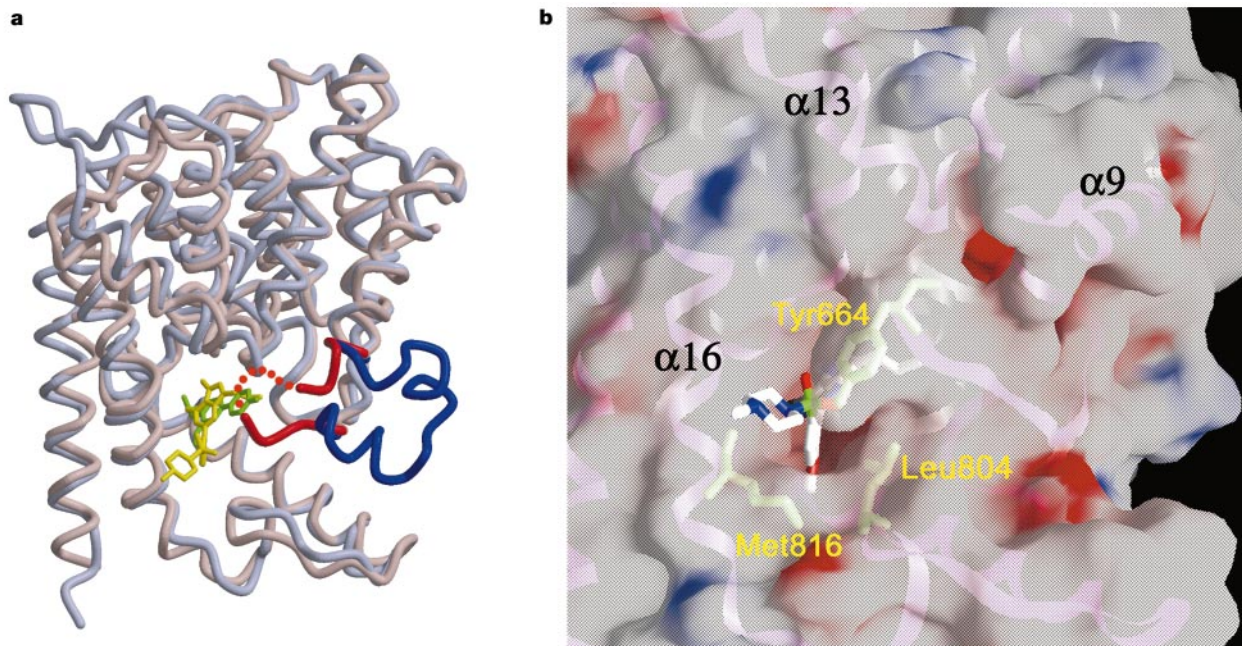


Figure 3 Comparison of PDE5 and PDE4 active sites. **a**, Superimposed C α traces of PDE5 (red) and PDE4 (blue)⁹ showing the difference between the two folds at the active site. Residues 304–325 of PDE5 and 660–680 of PDE4D are shown with deeper colours to emphasize the differences in this region. The stick model of sildenafil is shown in yellow

and that of zardaverine⁹ in green. **b**, Surface representation of active site pocket of PDE5. The molecular surface is coloured according to electrostatic potential (negative and positive in red and blue, respectively). Residues that form the active site pocket are shown in green. The bound sildenafil in PDE5 is shown as a stick model.

explain in part the weak binding of sildenafil to PDE4.

The L region composed of residues Tyr 664, Met 816, Ala 823 and Gly 819 surrounds the methylpiperazine group of sildenafil. Tyr 664 in PDE5, which corresponds to Phe 308 of helix α 8 in PDE4D, sits on an extended loop reaching towards the active site, so that residues 662–664 make a 'lid' over the pocket (Fig. 3). In contrast to the wide opening of the catalytic pocket in PDE4, the L region narrows the entrance to the active site of PDE5. This seems to be a unique structural feature of PDE5, which is retained in the PDE5–tadalafil complex, even though tadalafil does not interact with Tyr 664. The methylpiperazine group is exposed at the protein surface through the opening to the active site. Its interactions with the surrounding hydrophobic residues, such as Tyr 664 and Met 816, are not found in the equivalent region of PDE4–ligand complexes (Fig. 3), presumably another reason for the selectivity of sildenafil for PDE5. Notably, the sulphonyl group, a strong hydrogen bond acceptor, is not involved in any hydrogen bonding.

Vardenafil (K_d of 0.6–0.7 nM; selectivity over PDE6, approximately 16-fold; ref. 15) shows a very similar binding mode to sildenafil. The ethylpiperazine group makes no more interactions with PDE5 than the methylpiperazine of sildenafil.

The binding mode of tadalafil is different from that of sildenafil (Fig. 1a; see also Supplementary Fig. 2b). Tadalafil makes no interaction with the L region of the protein. The Q pocket also makes different interactions with the two ligands. The side chain of Gln 817 forms a single, not bidentate, hydrogen bond with the tadalafil NH group. In this respect, sildenafil probably resembles the natural cGMP substrate more closely than tadalafil. The H pocket, which accommodates the ethoxy group of sildenafil, is filled with the methylenedioxyphenyl group of tadalafil. The more extensive interactions with the H pocket may be one of the reasons that tadalafil maintains a high affinity without binding to the L region. Comparison of the structures of the three complexes suggests obvious modification (for example, changing the ethoxy group of sildenafil to fit to the H pocket) of the inhibitors to improve their binding affinity and selectivity.

Mutating the residues Tyr 612, His 653, His 657, His 682, His 684, His 685, Thr 723, Asp 764, Glu 785, Gln 789 or Glu 793 causes significant changes in cGMP binding^{14,16}. In our structural analysis, Tyr 612 interacts with the pyrazolopyrimidinone moiety of sildenafil. His 653 and Asp 764 are involved in zinc coordination. His 682 and His 685 coordinate the magnesium ion directly through water molecules. His 657 and Thr 723 are near the magnesium-binding site, and may have a role in positioning the substrate for efficient catalysis⁸. Glu 785, Gln 789, Glu 793 and His 684 seem to stabilize the overall structure. When we analyse a model of cGMP based on the structure of bound sildenafil in the PDE5 active site, a cyclic phosphate group binds to zinc and magnesium, replacing the bound water molecules bridging sildenafil to the metals (Supplementary Fig. 3).

Here we have presented the three-dimensional structures of PDE5 complexed with three inhibitors in clinical use: sildenafil, vardenafil and tadalafil. Our study provides a clear picture of the different binding modes of these ligands in the active site of PDE5. We believe that these structural studies will assist the discovery of potent and selective PDE inhibitors with improved pharmacological profiles. □

Methods

Purification, crystallization and data collection

The catalytic domain of PDE5A1 (residues Thr 537 to Gln 860; GenBank accession number U16545) amplified from a human muscle complementary DNA library was cloned into a T7 promoter-driven expression vector. The protein was overexpressed in *Escherichia coli* as a histidine-tag fusion protein, and purified by nickel and heparin affinity chromatography. The histidine tag was removed by thrombin cleavage. Specific activity of the purified protein was 2.5 nmol min⁻¹ mg⁻¹. Preparation of the selenomethionyl protein for MAD experiments was performed as described¹⁷. Sildenafil citrate (1-[[3-(6,7-dihydro-1-methyl-7-oxo-3-propyl-1H-pyrazolo[4,3-d]pyrimidin-5-yl)-4-

ethoxyphenyl]sulphonyl]-4-methylpiperazine citrate) was obtained from Z. No.

Tadalafil (6-benzo[1,3]dioxol-5-yl-2-methyl-2,3,6,7,12,12a-hexahydro-pyrazino[1',2':1,6]pyrido[3,4-b]indole-1,4-dione) was synthesized by the method of ref. 18. Vardenafil (2-[2-ethoxy-5-(4-ethylpiperazine-1-sulphonyl)-phenyl]-5-methyl-7-propyl-3H-imidazo[5,1-f][1,2,4]triazin-4-one) was obtained from D.-K. Kim. PDE5 and inhibitor were mixed at a molar ratio of 1:3 and the resulting complex was further purified by size-exclusion chromatography on a Superdex S200 column, and concentrated to 10 mg ml⁻¹ in the buffer (25 mM Tris-HCl, pH 7.5, 100 mM NaCl, 5 mM dithiothreitol) by ultra-filtration. The PDE5–inhibitor complex crystals were grown at 277 K by hanging-drop vapour diffusion, adding 1 μ l protein solution to 1 μ l well solution (100 mM Tris-HCl (pH 7.2–7.5), 0.2 M MgCl₂, 6–9% polyethylene glycol 8000). X-ray diffraction data were collected at 100 K either on 6B beamline at Pohang Light Source (PLS) or on BL44B2 beam line at the SPring-8 Synchrotron facility. In the case of the sildenafil and vardenafil complexes, crystals belong to space group C222₁, with unit cell parameters $a = 60.1$ Å, $b = 155.6$ Å and $c = 89.9$ Å. The PDE5–tadalafil complex crystals belong to space group C2, with $a = 131.4$ Å, $b = 48.6$ Å, $c = 123.9$ Å and $\beta = 117.3^\circ$. Although the PDE5–tadalafil complex data set has low data completeness (79.7%) owing to high mosaicity, the final electron density map is nevertheless of sufficient quality to position the ligand reliably. Diffraction data were processed and scaled using the HKL software package¹⁹. Supplementary Table 1 summarizes the statistics for data collection and refinement.

Structure determination and refinement

Eight of nine expected selenium sites in the asymmetric unit of the sildenafil-complexed crystal were identified with SOLVE²⁰. The phases were improved by solvent modification with RESOLVE²¹ to a figure of merit of 0.55 (at 2.8 Å). The resulting experimental map was of sufficient quality to place most residues of PDE5, except the 11 residues of the disordered region (residues 665–675). Sildenafil was identified at this stage. The model was built using the program O²² and refined with CNS²³. Subsequently, this refined Se-Met-substituted PDE5 model was used to solve the structure of native protein. Subsequent rounds of model adjustment, simulated annealing and thermal factor refinement were performed. During the final refinement stages, water molecules were inserted into the protein model. The final model contained residues 537–664, 676–830, one zinc ion, one magnesium ion, one sildenafil citrate molecule and 78 water molecules in the asymmetric unit, with 92.5% of residues in the most favoured regions of the Ramachandran plot and no residues in the disallowed region. The model has r.m.s. deviations of 0.008 Å in bond length and of 1.19° in bond angles. The R_{cryst} and R_{free} values of the final model were 19.5% and 24.5%, respectively (Supplementary Table 1).

The PDE5–tadalafil and PDE5–vardenafil structures were solved by molecular replacement using the structure of the PDE5–sildenafil complex as a search model. Both molecular replacement and model refinement were carried out using CNS. The final model of the tadalafil-complexed crystal contained two PDE5 catalytic domains (residues 537–664, 676–860), two zinc ions, two magnesium ions and two tadalafil molecules in the asymmetric unit. The final model of the vardenafil-complexed crystal contained one PDE5 catalytic domain (residues 537–664, 676–860), one zinc ion, one magnesium ion, one vardenafil molecule and 116 water molecules in the asymmetric unit.

Received 7 May; accepted 14 July 2003; doi:10.1038/nature01914.

- Beavo, J. A. Cyclic nucleotide phosphodiesterases: functional implications of multiple isoforms. *Physiol. Rev.* **75**, 725–748 (1995).
- Soderling, S. H. & Beavo, J. A. Regulation of cAMP and cGMP signaling: new phosphodiesterases and new functions. *Curr. Opin. Cell Biol.* **12**, 174–179 (2000).
- Corbin, J. D. & Francis, S. H. Cyclic GMP phosphodiesterase-5: target of sildenafil. *J. Biol. Chem.* **274**, 13729–13732 (1999).
- Rotella, D. P. Phosphodiesterase 5 inhibitors: current status and potential applications. *Nature Rev. Drug Discov.* **1**, 674–682 (2002).
- Conti, M., Nemoz, G., Sette, C. & Vicini, E. Recent progress in understanding the hormonal regulation of phosphodiesterases. *Endocr. Rev.* **16**, 370–389 (1995).
- Torphy, T. J. Phosphodiesterase isozymes. *Am. J. Respir. Crit. Care Med.* **157**, 351–370 (1998).
- Mehats, C., Andersen, C. B., Filopanti, M., Jin, S. L. & Conti, M. Cyclic nucleotide phosphodiesterases and their role in endocrine cell signaling. *Trends Endocr. Met.* **13**, 29–35 (2002).
- Xu, R. X. *et al.* Atomic structure of PDE4: Insights into phosphodiesterase mechanism and specificity. *Science* **288**, 1822–1825 (2000).
- Lee, M. E., Markowitz, J., Lee, J.-O. & Lee, H. Crystal structure of phosphodiesterase 4D and inhibitor complex. *FEBS Lett.* **530**, 53–58 (2002).
- Rybalkin, S. D., Rybalkina, I. G., Shimizu-Albergine, M., Tang, X. B. & Beavo, J. A. PDE5 is converted to an activated state upon cGMP binding to the GAF A domain. *EMBO J.* **23**, 469–478 (2003).
- Corbin, J. D. & Francis, S. H. Pharmacology of phosphodiesterase-5 inhibitors. *Int. J. Clin. Pract.* **56**, 453–459 (2002).
- Jeffrey, P. D. *et al.* Mechanism of CDK activation revealed by the structure of a cyclinA–CDK2 complex. *Nature* **376**, 313–320 (1995).
- Nikolov, D. B. *et al.* Crystal structure of a TFIIB–TBP–TATA-element ternary complex. *Nature* **377**, 119–128 (1995).
- Turko, I. V., Francis, S. H. & Corbin, J. D. Potential roles of conserved amino acids in the catalytic domain of the cGMP-binding cGMP-specific phosphodiesterase (PDE5). *J. Biol. Chem.* **273**, 6460–6466 (1998).
- Young, J. M. Expert opinion: Vardenafil. *Expert Opin. Invest. Drugs* **11**, 1487–1496 (2002).
- Sekhar, K. R., Gronin, P., Francis, S. H. & Corbin, J. D. *Phosphodiesterase Inhibitors* (eds Schudt, C., Dent, G. & Rabe, K. F.) 135–146 (Academic, New York, 1996).
- Double, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276**, 523–529 (1997).
- Orme, M. W., Sawyer, J. S. & Schultze, L. M. Indole derivatives as PDE5-inhibitors. *PCT Int. Appl. WO0236593 A1* (2002).

19. Otwinowski, M. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
20. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
21. Terwilliger, T. C. Maximum likelihood density modification. *Acta Crystallogr. D* **56**, 965–972 (2000).
22. Jones, T., Zou, J.-Y., Cowan, S. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
23. Brünger, A. T. *et al.* Crystallography & N.M.R. system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to J. R. H. Tame for a critical reading of the manuscript. We thank Z. No for providing sildenafil citrate; D.-K. Kim for providing vardenafil; D. K. Shin for discussion and figures; and H.-S. Lee and G.-H. Kim for their assistance at the Pohang Light Source (PLS), beamline 6B. Experiments at PLS were supported, in part, by the Ministry of Science and Technology (MOST) of Korea and POSCO. We also thank S.Y.P.'s group for their assistance at Spring-8 for high-resolution data. This work was supported partially by a grant from the National Research Laboratory Program and the Center for Biological Modulators of the 21c Frontier R&D Program, subsidized MOST. This work was also supported partly by Yuyu Inc. and KT&G Co. Ltd..

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.G.R. (sgro@crystalgenomics.com) or J.M.C. (jmcho@crystalgenomics.com). Coordinates for the sildenafil, tadalafil and vardenafil complex structure have been deposited in the Protein Data Bank under accession codes 1UDT, 1UDU and 1UHO, respectively.

erratum

The expression domain of *PHANTASTICA* determines leaflet placement in compound leaves

Minsung Kim, Sheila McCormick, Marja Timmermans & Neelima Sinha

Nature **424**, 438–443 (2003).

In the Methods section on page 443 of this Letter, a line was omitted. The sentence should read: “We determined antibody purity and specificity by western blot analysis on protein extracts prepared from wild-type and rs2 mutant apices.” □

corrigenda

Structure of the replicative helicase of the oncoprotein SV40 large tumour antigen

Dawei Li, Rui Zhao, Wayne Lilyestrom, Dahai Gai, Rongguang Zhang, James A. DeCaprio, Ellen Fanning, Andrzej Joachimiak, Gerda Szakonyi & Xiaojiang S. Chen

Nature **423**, 512–518 (2003).

The name of A. J. was misspelt in the author list and should be Andrzej Joachimiak. Also, he is in the Biosciences Division of SBC (and not at the Advanced Photon Source, as published). □

Impact of urbanization and land-use change on climate

E. Kalnay & M. Cai

Nature **423**, 528–531 (2003).

When calculating areal averages, which involve weighting gridded data with the cosine of the latitude of half-degree grid boxes, we divided the sum by the total number of grids but omitted to divide the sum also by the average cosine latitude of the domain, which is 0.786. This error affects only the area-averaged values, not the maps or the station values, or the relative differences between station and reanalysis values. As a result, the average numbers on the maps should be divided by this factor. As the proportions quoted remain the same, this error does not affect our conclusions, except that the values of area-averaged trends have to be multiplied by 1.272. The corrected estimate of the trend in daily mean temperature due to land use changes is 0.35 °C per century. □

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Global Head of Naturejobs:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)
**Netherlands/ Italy/ Spain/
Portugal/ Belgium:**
Evelina Rubio Hakansson (4973)
Scandinavia: Sille Opstrup (4994)
France/ Switzerland:
Amelie Pequignot (4974)
Natureevents:
Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakutamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Rental disagreements

Assuming that postdoc stipends are relatively equal (or, at least, equally low), housing allowances or rent subsidies can make a huge difference to a postdoc's quality of life — especially if they're doing their fellowship in an expensive urban area (see *Nature* **421**, 766–767; 2003). So discussions of proposals in two corners of the world to scale back or even eliminate housing subsidies is provoking concern among postdocs and faculty members alike.

At New York's Rockefeller University, the administration is considering plans to slowly reduce the rent subsidy it pays out until the grant is eliminated in 2007, and to subject postdocs living in university housing to an annual rent increase of up to 3%. These moves would, in effect, put the wages of Rockefeller's postdocs below the national average. Already, more than a third of its postdocs have complained to their representatives about the plans. Meanwhile, across the globe in Singapore, the government is reviewing the current policy of subsidizing postdoc housing by as much as 90% because rents have slipped during the recent recession.

One can be sympathetic, to a certain extent, to the economic pressures driving these deliberations. The stock market is down, deficits are up and money is tight. But scaling back subsidies now could have repercussions later. Rockefeller should still be able to attract postdocs, but it could find itself losing the cream of the crop to rival institutions. And the mere talk of reducing funds in Singapore might make it harder to attract world-class fellows to the island, at a time when the country is stepping up its recruiting efforts to match its infrastructure investments.

Should other institutions follow these leads, then the consequences for postdocs would be dire. But that is unlikely in a market that seeks the best talent — more likely, these few cuts will open the way to a more mixed postdoc pay system.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

Canada invests
in the future

p104

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS



Investing to compete Ontario

Aiming high: Ontario is ploughing money into its science and technology base, funding a wide range of projects including the intelligent transportation system (above) in Toronto.

Canada, unlike its big, brash neighbour to the south, is known for its modesty. So this spring, when the provincial government of Ontario announced plans to pour Can\$1 billion (US\$717 million) over the next ten years into a new cancer research institute, Canadian scientists and citizens alike took notice.

Calvin Stiller called the scale of the investment — Can\$100 for every person in the province — so audacious as to be “un-Canadian”. Not that Stiller is complaining. A physician who co-founded the healthcare corporation Diversicare and who has served on the executive committee of the Medical Research Council of Canada, Stiller says that the investment will invigorate the province. “Ontario is doing a very large

social-scientific economic experiment that I think will be seen ten years down the road as having been transformational,” he says.

That transformation will come not just from the cancer centre, but from a series of investments that the province began making just a few years ago. These include funds for research infrastructure, for recruiting and retaining scientific talent and for commercialization. The investments aim to turn the province into a world-class life-science region that will not only provide jobs for Canadian scientists but also make Ontario a more attractive destination for those from abroad.

About 100 new posts for senior scientists are expected, around half of which will be in the new cancer institute (surgeons, clinical investigators and

Lots of trust

The Ontario government’s massive investments in R&D back a wide variety of projects ranging from genomics and geophysics to the social sciences.

The Ontario Innovation Trust, for example, provides peer-reviewed funds to improve infrastructure in universities, colleges and hospitals. In its latest annual report, it lists 294 projects involving contributions from more than 300 industry partners to 31 March 2001. The funds are divided between health sciences (55.21%), natural sciences and engineering (37.47%),

multidisciplinary projects (6.70%), humanities and social sciences (0.57%), and arts and letters (0.05%).

The Ontario Research and Development Challenge Fund, designed to help attract and keep scientists in the province, provided Can\$435 million (US\$312 million) over four years and leveraged more than Can\$1.1 billion through private- and public-sector partnerships. Projects included the creation of new diagnostic and treatment tools for breast-cancer patients; two-way

fibre-optic connections for real-time data exchange; computational research using superconducting clusters; design of earthquake warning systems; and support for quantum-computing research.

The fund’s criteria are how a project will enhance the province’s capacity for research excellence; the performance and productivity of the institution and its leaders, researchers, students and business partners; and the potential jobs and economic growth that will result from the increased R&D activity.

D.S.

epidemiologists, among others). An as-yet-unknown number of investigators from other countries will also take part. A centre to bring together research, technology and commercialization — already started in downtown Toronto — will eventually create an estimated 6,000 full-time positions.

These investments began, oddly enough, with a strategy to privatize the province's hydroelectric regime. That led to public and private decisions to diversify the economy and bolster it by investing in the science and technology infrastructure, providing educational opportunities and creating an innovation culture.

The first manifestation of this philosophy is the Ontario Innovation Trust, which was set up in 1999 with a fund of Can\$250 million. The funding has since been expanded to more than Can\$1 billion to help the province's universities, hospitals, colleges and research institutes enhance their research and technology-development infrastructure. One of its awards, for example, helped Janet Rossant, head of development and fetal health at the Samuel Lunenfeld Research Institute, to set up a transgenic-mouse facility there and at the University of Toronto. The Ontario Innovation Trust's funds are matched by federal funds from the Canada Foundation for Innovation, and it operates at arm's length from the Ontario government.

A second instrument, the Ontario Research and Development Challenge Fund (of which Stiller is the chair), was then set up to help keep the best and brightest scientists, researchers and technicians in the province and to attract top researchers from elsewhere. Its Can\$750-million budget deals with what Ken Knox, formerly a deputy minister in the government and now president of the Innovation Institute of Ontario, calls "transforming research" such as robotics.

"We've got a significant platform in genomic research, stem-cell research and imaging," Knox says. The fund makes it possible for, say, an institution that has a core competency in imaging but is not strong enough to compete globally, to attract two more principal investigators and their technicians.

GROWTH INDUSTRY

Last year, Mick Bhatia's work with stem cells at the Robarts Research Institute in London, Ontario, received a Can\$5-million grant from the fund. "We wanted to test the physiological ability of stem cells to repair tissue," he says. "The approach we're taking is very pragmatic and very direct: simply to ask 'can these cells be placed to functionally replace damaged tissue?' Until now, that question has not been asked directly. Now we hope we can train individuals that are like-minded to think about that specific question."

To help institutions bring in young scientists, a Premier's Research Excellence Awards Program was established in 1998, with a pot of Can\$75 million. It

provides young scientists with Can\$100,000 to hire technicians, allowing them to complete their work faster — something that potential contenders said would encourage them to stay in Ontario. Finally, the province set up an annual Platinum Award of Can\$1 million to each of its two "most outstanding scientists" in whatever field. Last year, just one was awarded: to Tony Pawson, scientific director of the Samuel Lunenfeld Research Institute in Toronto.

To promote commercialization of research, the government has now begun construction of an area called MaRS: the discovery district for Medical and Related Sciences. "We said we've got all this money going into research and we're not getting

commercialization results," says Knox. "It's because scientists want to be able to walk between their lecture theatre, their laboratory and their place of business, as they can do in Boston. The future is the convergence of all the sciences."

MaRS will be located next door to the University of Toronto, one of the largest universities in North America, and within walking distance of Ryerson University, seven internationally renowned hospitals, and more than 30 research institutes and centres in downtown Toronto. It will include all of the components necessary for commercialization — industrial firms, venture capital, accountants, lawyers and government funding offices — all under a single roof.

The idea is to provide all the services necessary to assist start-up companies through the stages of commercialization. Construction for the first three buildings, which will total about 60,000 square metres, started in October 2002. Phase one is expected to cost about Can\$165 million and the second phase about Can\$180 million, with tenants occupying the buildings in the fourth quarter of 2004.

Total commitments of Can\$65 million have been received from federal and provincial governments, the city of Toronto, and private individuals and companies. Stiller calls these efforts a way to compete with the two 'benchmarks' of Massachusetts and California.

"The Ontario Cancer Research Institute is even a further leap," Stiller adds, referring to another new venture. Initial reaction to the cancer-institute announcement has been positive. Mark Lievonen, president and chief executive of Aventis Pasteur, Toronto, says: "We're very dependent upon the R&D skills that we can bring and the medical/clinical structure that's available, and something like this is very exciting." He believes that private-sector jobs will result from a partnership with the new provincial centre.

Pawson says that the new cancer centre will bring Ontario to "a new level", in terms of integrating basic research with clinical practice. And long-term research investment may become a Canadian characteristic worth making some noise about.

David Spurgeon is a freelance science writer based near Montreal.

♦ www.oit.on.ca

♦ www.ontariochallengefund.com



Ontario's premier, Ernie Eves, discusses the province's technology investments at the University of Toronto.